UNIVERSIDADE DE BRASÍLIA - UnB
FACULDADE DE CIÊNCIA DA INFORMAÇÃO - FCI
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO – PPGCINF

HENRIQUE MONTEIRO CRISTOVÃO

UM MODELO HÍBRIDO DE RECUPERAÇÃO DE INFORMAÇÃO E CONHECIMENTO BASEADO NA SÍNTESE DE MAPAS CONCEITUAIS OBTIDOS POR OPERAÇÕES DE TRANSFORMAÇÃO DE REDES COMPLEXAS ORIENTADAS POR BUSCA DE RELACIONAMENTOS ENTRE TERMOS DE CONSULTA EM BASES DE DADOS LIGADOS

BRASÍLIA
2016

HENRIQUE MONTEIRO CRISTOVÃO

UM MODELO HÍBRIDO DE RECUPERAÇÃO DE INFORMAÇÃO E CONHECIMENTO BASEADO NA SÍNTESE DE MAPAS CONCEITUAIS OBTIDOS POR OPERAÇÕES DE TRANSFORMAÇÃO DE REDES COMPLEXAS ORIENTADAS POR BUSCA DE RELACIONAMENTOS ENTRE TERMOS DE CONSULTA EM BASES DE DADOS LIGADOS

Tese apresentada ao Programa de Pós-Graduação em Ciência da Informação da Faculdade de Ciência da Informação da Universidade de Brasília, como requisito parcial à obtenção do título de Doutor em Ciência da Informação.

Área de concentração: Gestão da Informação

Linha de pesquisa: Organização da Informação

Orientador: Prof. Dr. Jorge Henrique Cabral Fernandes

BRASÍLIA
2016

Cristovão, Henrique Monteiro
CC933m Um modelo híbrido de recuperação de informação e conhecimento baseado na síntese de mapas conceituais obtidos por operações de transformação de redes complexas orientadas por busca de relacionamentos entre termos de consulta em bases de dados ligados / Henrique Monteiro Cristovão; orientador Jorge Henrique Cabral Fernandes. -- Brasília, 2016.
315 p.

Tese (Doutorado - Doutorado em Ciência da Informação) -- Universidade de Brasília, 2016.

1. Recuperação de Informação e Conhecimento. 2. Análise de Redes Complexas. 3. Mapas Conceituais. 4. Dados Abertos Ligados. 5. Web Semântica. I. Fernandes, Jorge Henrique Cabral, orient. II. Título.

FOLHA DE APROVAÇÃO

Título: "Um Modelo Híbrido de Recuperação de Informação e Conhecimento Baseado na Síntese de Mapas Conceituais Obtidos por Operações de Transformação de Redes Complexas Orientadas por Busca de Relacionamentos entre Termos de Consulta em Bases de Dados Ligados"

Autor (a): Henrique Monteiro Cristovão

Área de concentração: Gestão da Informação

Linha de pesquisa: Comunicação e Mediação da Informação.

Tese submetida à Comissão Examinadora designada pelo Colegiado do Programa de Pós-graduação em Ciência da Informação da Faculdade em Ciência da Informação da Universidade de Brasília como requisito parcial para obtenção do título de **Doutor** em Ciência da Informação.

Tese aprovada em: 12 de agosto de 2016.

Prof. Dr. Jorge Henrique Cabral Fernandes
Presidente (PPGCINF/UnB)

Prof. Dr. Rogério Henrique de Araújo
Membro Interno (PPGCINF/UnB)

Prof. Dr. Ricardo Barros Sampaio
Membro Interno (PPGCINF/UnB)

Prof. Dr. Alberto J. Cañas
Membro Externo (IHMC)

Profª Drª Fernanda Lima
Membro Externo (CIC/UnB)

Prof. Dr. Fernando Willian Cruz
Suplente (PPGCINF/UnB)

Dedico essa tese à minha família.

AGRADECIMENTOS

Ao meu orientador, prof. Jorge Henrique Cabral Fernandes, pelas orientações precisas e por incentivar ampla liberdade nas investigações e busca de melhores caminhos no desenvolvimento do trabalho.

Ao meu orientador do estágio de pesquisa (doutorado sanduíche), prof. Alberto J. Cañas, realizado no Florida Institute for Human & Machine Cognition (IHMC)/EUA, pela sabedoria, disponibilidade e valiosas orientações.

A todos que se envolveram e colaboraram com a criação, aprovação e condução do Doutorado Interinstitucional (Dinter) em Ciência da Informação, tanto no âmbito da Universidade Federal do Espírito Santo (UFES) quanto da Universidade de Brasília (UnB), em especial à prof.ª Dulcineia Sarmento Rosemberg, pela sua dedicação em todo o processo.

Aos professores membros da banca de qualificação, prof.ª Fernanda Lima, prof.ª Ivette Kafure Muñoz, por suas valiosas contribuições em direção da melhoria do trabalho.

Aos professores membros da banca de defesa, prof. Alberto J. Cañas, prof.ª Fernanda Lima, prof. Rogério Henrique de Araújo Junior, prof. Ricardo Barros Sampaio e prof. Fernando William Cruz, pela disponibilidade em avaliar e contribuir com o seu conhecimento nessa etapa final.

Ao Departamento de Computação e Eletrônica/UFES, ao qual sou lotado como professor, por não medir esforços quanto à viabilização do meu afastamento das atividades docentes em parte do doutorado.

Ao Florida Institute for Human & Machine Cognition (IHMC)/EUA, pela oportunidade de passar nove meses em contato com pesquisadores, em especial Larry Bunch por suas valiosas contribuições.

A CAPES e ao CNPq pelo apoio financeiro em determinados momentos do doutorado.

Ao Brasil por ter me fornecido educação pública, gratuita e de qualidade, usufruída por mim no ensino fundamental e médio, graduação, mestrado e doutorado.

O mundo não é, está sendo.

Paulo Freire

Internet permite la creación en red, más allá de una suma de individualidades.

Manuel Castells

Imagine a world in which every single human being can freely share in the sum of all knowledge.

Wikimedia Foundation

RESUMO

Investiga e desenvolve um modelo de recuperação de informação e conhecimento em bases de dados ligados, baseado na transformação de redes complexas obtidas a partir da busca por relacionamentos entre os termos de uma consulta formulada pelo usuário. Os relacionamentos derivados dos termos da consulta são obtidos a partir de buscas em uma base de dados ligados (*linked data*), gerando uma cadeia de transformações sobre redes de triplas RDF, baseadas em métricas de análise de redes complexas. A pesquisa é exploratória, buscando de forma simultânea o desenvolvimento de formulações teóricas sobre sistemas de recuperação de informação, o estudo de suas implicações práticas do ponto de vista das tecnologias atualmente disponíveis, e a coleta de dados empíricos decorrentes de seguidas iterações de modelagem, implementação e teste de protótipos de um sistema computacional. Operações de análise de redes complexas foram usadas para ranqueamento e seleção de informações recuperadas, que depois retroalimentam novas buscas para expansão da rede de informação. Por intermédio do protótipo implementado, o modelo foi validado com um grupo de dezessete usuários. O grau de satisfação com os resultados sugere qualitativamente que o modelo possui boa precisão e revocação. Foram de grande relevância as iteradas operações de transformação das redes na retroalimentação do modelo. Há indícios da necessidade de maiores bases de conhecimento. A rede final é mapeada em um mapa conceitual, e o modelo desenvolvido foi analisado à luz da equação fundamental da Ciência da Informação de Brookes e da aprendizagem significativa de Ausubel. A natureza interdisciplinar do modelo permitiu explorar de forma diferenciada e original a recuperação de informação e conhecimento em bases de dados ligados, apontando relevância de investigações futuras. O resultado culminou no desenvolvimento de um modelo híbrido de recuperação de informação e conhecimento que é simultaneamente fundamentado nos três paradigmas correntes da Ciência da Informação: físico, cognitivo e social.

Palavras-chave: Ciência da Informação. Recuperação de Informação. Recuperação de Conhecimento. Web Semântica. Dados Abertos Ligados. Ciência das Redes. Redes Complexas. Análise de Redes Complexas. Mapas conceituais. Aprendizagem Significativa.

ABSTRACT

It researches and develops a model of information and knowledge retrieval on bases of linked data, based on the transformation of complex networks obtained from the search for relationships about the terms of a user-formulated query. The relationships derived from the query terms are obtained from a search in a base of linked data. It generates a sequence of transformations on RDF triples networks, based on metrics of complex networks analysis. The research is exploratory and it was developed from simultaneous way to theoretical formulations on information retrieval systems, the study of its practical implications in point of view of the currently available technologies, and the collection of empirical data resulting from several iterations of modeling, implementation and prototype testing of a computer system. Complex networks analysis operations were used for ranking and selection of retrieved information, which then uses a bootstrapping process to feedback the new search to the expansion of the information network. The model was validated over the implemented prototype, with a group of seventeen users. The degree of satisfaction with the results suggests, qualitatively, that the model has good precision and recall. The iterated operations of network transformation were of great importance in the feedback process of the model. There is evidence of the need for larger knowledge bases. The final network is mapped to a conceptual map, and the model developed was analyzed in light of the fundamental equation of the Information Science of the Brookes and the meaningful learning of the Ausubel. The interdisciplinary nature of the model allowed us to explore the different and original way information and knowledge retrieval on bases of linked data, indicating relevance of future investigations. The results culminated in the development of a hybrid model of information and knowledge retrieval that is grounded simultaneously in the three current paradigms of information science: physical, cognitive and social.

Keywords: Information Science. Information Retrieval. Knowledge Retrieval. Semantic Web. Linked Open Data. Science Networks. Complex networks. Complex networks analysis. Concept Maps. Meaningful Learning.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ARS	Análise de Redes Sociais
ASCII	*American Standard Code for Information Interchange*
CI	Ciência da Informação
CIR	*Cognitive information retrieval*
CSV	*Comma-separated values*
DRS	*Data retrieval systems*
GEXF	*Graph Exchange XML Format*
HTML	*HyperText Markup Language*
IHMC	Florida Institute for Human & Machine Cognition
IIR	*Interactive information retrieval*
IRI	*Internationalized Resource Identifier*
JSON	*JavaScript Object Notation*
KOS	*Knowledge Organization System*
LOD	*Linked Open Data*
LOGD	*Linked Open Government Data*
OGD	*Open Government Data*
OWL	*Web Ontology Language*
PDF	*Portable Document Format*
RAD	*Rapid Application Development*
RDF	*Resource Description Framework*
RI	Recuperação de informação
RIF	*Rule Interchange Format*
SKOS	*Simple Knowledge Organization System*
SNA	*Social Network Analysis*
SPARQL	*SPARQL Protocol and RDF Query Language*
SQL	*Structured Query Language*
URI	*Internationalized Resource Identifier*
URL	*Uniform Resource Locator*
URN	*Uniform Resource Name*
W3C	*World Wide Web Consortium*
web	*World Wide Web*
WWW	*World Wide Web*

SUMÁRIO

1 INTRODUÇÃO

As desigualdades de riqueza, poder e oportunidades na sociedade são, por consequência, refletidos como desigualdades no acesso à informação (VICKERY; VICKERY, 1987). Existem muitas variáveis que determinam o acesso à informação, tal como o comportamento do usuário (SARACEVIC, 2010) e suas necessidades informacionais que, juntamente com a sua organização em processos de comunicação, é um dos focos da Ciência da Informação (CI) (WERSIG; NEVELING, 1975). O estudo dessas necessidades informacionais do usuário, que precede a CI, tem tomado um novo rumo desde o surgimento da World Wide Web (web), que estabeleceu novas tecnologias de organização, busca e disseminação da informação (SARACEVIC, 2010). Nesse contexto, Stuckenschmidt (2012) afirma que a web é uma área de pesquisa inequívoca, pois possui desafios muito variados. A web é um fenômeno social de grande escala que produziu propriedades emergentes e comportamentos transformadores (SHADBOLT *et al.*, 2013). Hendler *et al.* (2008) chamaram de Ciência da Web o estudo interdisciplinar em torno da web. Shadbolt *et al.* afirmaram que a Ciência da Web, muito próxima das Ciências Humanas e Sociais, é reconhecida como uma importante área de estudo, composta de um corpo organizado de conhecimento, e que irá reunir uma nova geração de mentes curiosas. Um campo grande de pesquisas, estudos, experimentos e descobertas têm emergido a partir da Ciência da Web e, paralelamente, muitos problemas apareceram desafiando usuários e pesquisadores da CI e de outras áreas do conhecimento.

Gleick (2011) remete a um paradoxo cada vez mais atual: o fato de existir uma quantidade muito grande de informações como jamais se viu não leva aos humanos sentirem-se mais inteligentes, mas, pelo contrário, há um sentimento de esmagamento devido à enxurrada de informações. De fato, não surpreende, portanto, a web ser uma bagunça e não se admira que todo mundo esteja interessado em alguma forma de recuperação de informação como uma solução para consertá-la (SARACEVIC, 1999). Esta área, a Recuperação de Informação, é um campo interdisciplinar de vários domínios, desde a CI até Ciência da Computação tendo várias ferramentas de apoio como: classificação, tesauros, taxonomia e ontologias (PONTES JUNIOR; CARVALHO; AZEVEDO, 2013). Apesar de ainda ser um tema controverso, devido a existência de várias opiniões, os autores sinalizam que a Recuperação de Informação, sob o ponto de vista da CI, é atualmente referida como Recuperação de Conhecimento.

Mapa conceitual pode ser considerado uma ferramenta importante para organizar e representar o conhecimento, dando suporte à aprendizagem significativa (NOVAK, 1977) e com comprovações acerca da sua eficácia na representação do conhecimento (CAÑAS *et al.*, 2005). A teoria da Aprendizagem Significativa de Ausubel (1968) estabelece que uma nova informação se relaciona com a estrutura de conhecimento de um indivíduo para formar novos conhecimentos. Os mapas conceituais podem ser universais e ubíquos (NOVAK; CAÑAS, 2010), ou seja, eles têm uma aplicabilidade muito ampla e podem estar presentes em atividades realizadas por usuários de todas as idades, em todas as partes do mundo e elaborados em qualquer linguagem, e ainda podem ser encontrados em muitos ambientes, tecnológicos ou não. O mapa conceitual pode ser usado para disseminar informações de uma forma mais eficiente. Vekiri (2002) mostra a importância e o papel das representações gráficas sobre os mecanismos cognitivos envolvidos na aprendizagem, principalmente quando eles estão combinados com texto, tal como acontece com os mapas conceituais.

Um dos maiores desafios na área da gestão da informação inteligente é a exploração da web como uma plataforma de integração de dados e informações, como também para pesquisa (AUER *et al.*, 2013). Considerada como uma extensão da web atual, a Web Semântica foca na atribuição de significados à informação, permitindo que computadores e pessoas trabalhem em conjunto no acesso a essa informação, pois, no momento, a maior parte do conteúdo da internet é destinada ao consumo humano e não compreensível por software. O funcionamento da Web Semântica confere à web atual a capacidade de agregar múltiplos dados relacionados entre si. Há mais de uma década Berners-Lee, Hendler e Lassila (2001) sinalizavam o quanto a Web Semântica poderia colaborar com a evolução do conhecimento humano como um todo, sendo que ela não é simplesmente uma ferramenta para realização de tarefas individuais. Os autores também destacaram que na medida em que o conhecimento humano fosse construído de forma colaborativa com a Web Semântica munida de ontologias e agentes de software, seria possível auxiliar a comunicação e o trabalho colaborativo entre diversos povos de culturas e/ou línguas diferentes. De fato, Stuckenschmidt *et al.* (2013) após análise de 10 anos de conferências sobre a Web Semântica, confirmam que essa área, tal como outros campos emergentes, passou por mudanças significativas no que diz respeito à importância e qualidade do trabalho experimental, ficando no caminho de se tornar uma disciplina científica estabelecida com altos padrões relativas à avaliação experimental dos trabalhos.

Junto a essa evolução está ocorrendo um grande movimento de abertura dos dados[1] por muitos governos e instituições, e um conjunto de melhores práticas para a publicação e ligação de dados estruturados na web conhecido, mundialmente, como *Linked Data*[2] e, na literatura nacional, como dados ligados. Navegadores específicos e motores de busca estão se adaptando aos dados ligados para funcionar sobre a Web Semântica oferecendo serviços diferenciados (BIZER; HEATH; BERNERS-LEE, 2009), pois em contraste com os bancos de dados clássicos, que exigem um investimento inicial e caro na modelagem, tecnologias de dados ligados permitem às empresas terem menos esforço e, assim, investirem em ligações, vocabulários compartilhados entre outros para permitir uma integração mais profunda (HEATH; BIZER, 2011). Quando os dados abertos ocorrem em contextos de dados abertos, dá-se o nome de *Linked Open Data* (LOD), ou dados abertos ligados. Eles possibilitam a criação de serviços diferenciados que facilitam a inovação e a criação de conhecimento a partir de dados interligados, sendo um mecanismo importante para gestão da integração da informação (BAUER; KALTENBÖCK, 2012). Por exemplo, Santos Neto *et al.* (2013) avaliaram o impacto de LOD na integração de dados de arquivos, bibliotecas e museus, que normalmente possuem acervos isolados com seus códigos próprios e maneiras particulares de representar a informação, e sugeriram aos profissionais da informação que usufruam da potencialidade de LOD buscando novas aplicações para ampliar a interoperabilidade dos dados disponíveis na web.

A Ciência das Redes, ou *Network Science*, é o conhecimento organizado de redes baseado em estudos realizados por métodos científicos (NATIONAL RESEARCH COUNCIL, 2005). É um campo de pesquisa interdisciplinar que atualmente faz contribuições significativas enquanto ferramenta e método no auxílio à resolução de problemas em diversas áreas. Sobre as contribuições da área, o sociólogo Manuel Castells (2000) já sinalizava que a investigação sobre a estrutura em rede da economia global ajudaria a se pensar em estratégias e políticas adequadas e sintonizadas com o nosso tempo. Segundo Newman (2010), uma variedade grande de problemas pode ser representada por redes. A Ciência das Redes, que também envolve a Web, é uma tentativa de compreender as redes emergentes na natureza, a tecnologia e a sociedade através de um conjunto de ferramentas e princípios (BARABÁSI, 2013). Nesse contexto, o conceito de redes complexas surgiu para denotar aquelas não

[1] Dados Abertos são dados que podem ser livremente utilizados, reutilizados e redistribuídos por qualquer um (OPEN DEFINITION, 2014). Ver a subseção 2.3.3 do referencial teórico.
[2] *Linked Data* é descrito e explicado em detalhes no site: <http://linkeddata.org/>

triviais e que representam a maioria das redes de nosso mundo em várias áreas do conhecimento, como por exemplo, as redes informacionais encontradas na web.

A análise de redes complexas é uma área de estudo que auxilia a compreensão de fenômenos de uma rede complexa, sendo mais abrangente do que a análise de redes sociais (ARS), onde essa última é mais referenciada na literatura científica. Para Wasserman e Faust (1994), ARS é o estudo aplicável a redes cujos nós são entidades sociais. Porém, esse campo de estudo pode ser aplicado a redes cujos nós tenham uma origem diferente de uma entidade social tal como ocorrem com as redes de informações (NEWMAN, 2010). Segundo esse autor, muitas das ideias importantes dessa área vieram das Ciências Sociais, e grande parte da linguagem usada para descrever essas ideias reflete origem sociológica, no entanto, os métodos descritos são agora amplamente utilizados em diversas áreas. As pesquisas em redes complexas estão cada vez mais revelando o quão as ideias estão conectadas, isto é, descobertas na Biologia, na Ciência da Computação, na Sociologia e na Física podem estar intimamente conectadas (BUCHANAN, 2002).

A natureza eminentemente interdisciplinar das áreas CI, Ciência das Redes, Ciência da Web e mapas conceituais fundamentados na Aprendizagem Significativa trazem possibilidades diferenciadas no atendimento às demandas da sociedade. A união dessas áreas revela pontos de interseção e também permite que elementos de uma área sejam visualizados ou interpretados sob a ótica de outra área, conseguindo, assim, potencializar soluções específicas que antes eram resolvidas num contexto meramente disciplinar.

1.1 Problema

Vannevar Bush, nos anos 40, já sinalizava que a capacidade humana em gerar novas informações era muito maior do que a sua capacidade em armazená-la de forma a permitir futuras consultas (BUSH, 1945). Consequentemente, ele estava alertando sobre a necessidade de melhorar a recuperação de informação e conhecimento. De fato, atualmente os usuários são desafiados com uma quantidade muito grande e crescente de dados na web que tem gerado uma necessidade de desenvolvimento de sistemas de recuperação de informação cada vez mais sofisticados (USBECK, 2014). Stuckenschmidt (2012) também destaca que um dos maiores desafios da gerência de dados semânticos da web é provavelmente a escalabilidade, ou seja, a capacidade de lidar com grande volume de dados.

A web está sendo ameaçada, alerta Berners-Lee (2010). Entre as várias indicações do autor, destaca-se a compartimentalização de informações por grandes provedores,

principalmente no ambiente de redes sociais. Bauer e Kaltenböck (2012) também alertam que a natureza especialista de muitas empresas reflete de forma negativa na forma como as informações são geridas, pois muitos dados não são ligados e muita informação está escondida. Um dos problemas associados a isso é a falta de definição de seus formatos e que está dentro de um contexto maior chamado de interoperabilidade, ou seja, a capacidade de diversos sistemas e organizações trabalharem juntos.

A interoperabilidade é um dos elementos centrais para a habilidade de combinar conjuntos de dados de diferentes fontes, possibilitando o desenvolvimento de melhores produtos e serviços e em maior quantidade (OPEN DATA HANDBOOK DOCUMENTATION, 2012). Na mesma linha, Ding, Peristeras e Hausenblas (2012) alertam que a integração dos dados em vários domínios e formatos, com vocabulários diferentes e ainda acompanhados de metadados de qualidade variável é um dos principais desafios para os dados governamentais abertos ligados, conhecidos na literatura mundial por *linking open government data*. Estes autores também destacam que, embora esses dados tenham sido a parte da Web Semântica que mais cresce, a maioria dos pesquisadores reconhece que existe uma barreira grande para iniciar a sua produção. Para minimizar o problema da baixa interoperabilidade, o World Wide Web Consortium[3] (W3C), principal órgão de padronização na área de tecnologias web, emite recomendações sobre as principais linguagens e protocolos (MIKA, 2007).

A visualização e consulta a dados abertos ligados ainda é limitada e carece de melhores formas para atender as necessidades do usuário, tanto de busca de informações quanto da interpretação das informações dos resultados das buscas. Não basta haver conexão entre as informações para que elas sejam acessíveis diretamente por usuários. Lima (2005) analisou vários estudos sobre a navegação em hipertextos e observou grande desorientação do usuário na navegação onde nem sempre ele consegue encontrar conscientemente a informação desejada.

Considerando os problemas apontados nessa seção, tais como, a falta de capacidade em lidar com grande volume de informações na web e Web Semântica, a necessidade em melhorar a recuperação de informação, o problema da compartimentalização de informações na web, a necessidade de abertura de dados, a demanda pela melhoria na interoperabilidade da web, a necessidade em melhorar a consulta e visualização de dados ligados, e ainda que os

[3] World Wide Web Consortium (W3C): é um consórcio internacional que agrega empresas, órgãos governamentais e organizações independentes com a finalidade de estabelecer padrões para a criação e a interpretação de conteúdos para a Web. Fundado por Tim Berners-Lee em 1994. Disponível em: <http://www.w3.org/>.

mapas conceituais podem melhorar significativamente a disseminação e leitura de informações, a partir de pesquisas como O'Donnel (1993), Vekiri (2002) e Orrantia (2012) e, também considerando o poder e aplicabilidade das técnicas de análise de redes complexas, a questão central desse projeto é: **como recuperar informação e conhecimento em bases de dados ligados com ênfase nos relacionamentos relevantes entre os termos de uma busca que represente uma necessidade informacional do usuário?**

1.2 Objetivos da pesquisa

1.2.1 Objetivo geral

Desenvolver um modelo para recuperação de informação e conhecimento no contexto dos dados ligados que revele relacionamentos entre termos de uma consulta associada à necessidade informacional do usuário, usando operações de manipulação de redes complexas e geração de mapas conceituais.

1.2.2 Objetivos específicos

a) Integrar conceitualmente os temas: Recuperação de Informação e Conhecimento, Redes Complexas, Mapas Conceituais, e Dados Abertos Ligados;
b) Investigar transformações baseadas na análise de redes complexas que podem ser empregadas na seleção e ranqueamento de relacionamentos em redes de informação, priorizando a descoberta de relacionamentos que satisfaçam uma necessidade informacional.
c) Elaborar um modelo para mapear o fluxo informacional entre dados ligados, redes de informação e mapas conceituais.
d) Desenvolver e validar junto a um grupo de usuários um protótipo executável computacional que represente o modelo desenvolvido.

1.3 Justificativa

O foco nas necessidades do usuário da informação tem aumentado ao longo da história em comparação a ênfase dada nos sistemas de informação e equipamentos, segundo Saracevic (2007), Wilson (2000) e Casado (1994). Desde os anos 80 também se observa que, com o advento das novas tecnologias na vida cotidiana das pessoas, o comportamento intelectual e

afetivo nos usuários tem mudado (BABIN; KOULOUMDJIAN, 1983). Em sintonia a essas preocupações no contexto do usuário, espera-se como resultado dessa pesquisa, que um contexto informacional em dados ligados possa ser mais bem disseminado e compreendido pelo usuário a partir da sua reapresentação por intermédio de mapas conceituais atendendo melhor suas necessidades informacionais e ficando a altura de seu tempo tecnológico.

Pontes Junior, Carvalho e Azevedo (2013) sugerem que os métodos de recuperação de informação (RI) poderiam ser mais flexíveis, e indicam a possibilidade da existência de relações entre os conceitos, usados numa taxonomia; os autores ainda refletem sobre um possível caminho onde essas relações poderiam ser estabelecidas pelo próprio usuário, ficando registrados numa rede. Assim, seria possível a "[...] representação de uma estrutura de conhecimentos confeccionada por múltiplos usuários, que, por sua vez, possibilitaria a detecção de conceitos centrais e amarrados [...]" (p. 15). O estudo da recuperação de informação tradicional foca na modelagem da relevância entre uma consulta textual e os documentos da base de conhecimento, e o ranking com os documentos recuperados não tem sido satisfatório, principalmente no contexto da web (LIU, 2006). A autora faz diversas indicações de estudos que usaram a análise de grafos como parte do processo da RI e conclui que isso pode ser uma boa saída para a melhoria da RI. Além disso, Zhang (2008) indica benefícios para o uso de uma RI mais visual, tal como a redução da carga cognitiva do usuário. Sintonizado a essas indicações, o modelo proposto trabalha com relacionamentos entre conceitos recuperados bem como o uso de análise em grafos e a apresentação visual da informação recuperada por intermédio de mapas conceituais.

A web precisa de ferramentas eficientes para gerenciar, recuperar e filtrar informações (BAEZA-YATES; RIBEIRO-NETO, 2011). Além disso, os modelos clássicos de recuperação de informação normalmente não se adaptam ao ambiente web, pois eles foram desenvolvidos para ambientes fechados onde todo o universo documental é restrito e pode ser conhecido pelo desenvolvedor (SILVA; SANTOS; FERNEDA, 2013). O presente trabalho sugere um modelo híbrido para recuperação de informação e conhecimento que vai além do convencional, usando redes complexas e mapas conceituais.

Lévy e Authier (1995), por intermédio de um processo denominado 'Árvores de Conhecimentos', sugerem a socialização dos saberes pertencentes a cada cidadão, e dessa forma democratizar o conhecimento estabelecendo formas de medição sobre os conhecimentos e competências de cada pessoa e motivando-as ao desenvolvimento de novas aprendizagens. Levy e Authier falavam de um processo que, mais recentemente, ganhou força com o movimento mundial dos dados abertos. O manual Open Data Handbook

Documentation (2012) exemplifica áreas e atividades em que os dados abertos estão gerando valor: transparência e controle democrático, participação, empoderamento, melhorias em produtos e serviços privados, melhoria da eficiência e da eficácia de serviços públicos, medição do impacto de políticas, e novos conhecimentos a partir da mescla de fontes de dados e padrões. De uma maneira geral, a captura e a representação da semântica de dados na web dará um enorme potencial para aplicações avançadas em uma área que ganha rapidamente importância em quase todas as áreas de negócio e da sociedade, incluindo o comércio eletrônico, o discurso político e intercâmbio científico, e sendo essa combinação um desafio de pesquisa de longo prazo, faz a semântica de dados na web um tópico que promete permanecer em relevância por muito tempo (STUCKENSCHMIDT, 2012). O modelo proposto no presente trabalho adota o modelo de dados abertos ligados como base de conhecimento e vem reforçar e motivar o crescimento dessas bases, principalmente, oferecendo um método de disseminação, valorização e utilização dos dados abertos ligados no formato de mapas conceituais.

Um dos aspectos mais interessantes dos dados abertos ligados é que eles permitem a sua visualização sob uma ótica diferenciada oportunizando a criação de serviços inovadores. Stuckenschmidt (2012) argumenta que a consequência mais interessante advinda da publicação de dados estruturados e sua semântica na web é a oportunidade de usar esses dados como base para a análise de problemas complexos, que exigem a observação combinada de dados de múltiplas fontes que antes eram difíceis ou mesmo impossíveis de acessar e combinar de uma forma significativa. Um dos princípios primários e fundamentais da web é a sua universalidade e, projetos como dados abertos ligados são importantes para manutenção dessa universalidade. Segundo Berners-Lee (2010), manter a web universal e seus padrões abertos ajudam as pessoas a inventar novos serviços. Por exemplo, a abertura de dados governamentais públicos, sem restrições legais, financeiras, tecnológicas ou políticas, permitirá uma redefinição do papel dos governos enquanto fornecedores de informação viabilizando a agregação de novos valores pelos cidadãos, com novas aplicações e, consequentemente, novos conhecimentos gerados (BAUER; KALTENBÖCK, 2012). De fato, a combinação de dados de variadas fontes pode criar conhecimentos e inspirar novas ideias, que vão criar novos campos de aplicação. O manual Open Data Handbook Documentation (2012) exemplifica situações criadas com ideias inesperadas que surgiram a partir da combinação de bases de dados abertas de diferentes locais. A visualização do resultado da busca, proposta pelo modelo do presente trabalho, traz um aspecto inovador quando informações advindas de bases de dados ligados são mapeadas em mapas conceituais.

Aplicações da Ciência das Redes estão em várias áreas do conhecimento humano e têm contribuído de forma significativa, como nos três seguintes exemplos: (i) Baker e Faulkner (1993) descobriram uma estrutura de redes ilegais associada ao crime organizado que formavam uma prática ilegal de fixação de preços nas vendas de equipamentos elétricos pesados; (ii) Goh *et al.* (2007) fizeram uma rede para indicar a origem genética de doenças, servindo de base a outras pesquisas na área da saúde; (iii) Mika (2007) relatou em seu livro o quanto as tecnologias associadas às redes e a Web Semântica auxiliaram na resolução de problemas advindos do furacão Katrina ocorrido nos EUA em 2005. Mika ainda argumenta que compreender o papel das redes na ciência é um primeiro passo importante para organizar o processo científico de maneira mais eficiente. Existe uma quantidade grande, na literatura, de problemas resolvidos por processos fundamentados em redes. De fato, “o conexionismo generalizado da sociedade das redes de computadores criou novas formas de espaço e tempo, um espaço e um tempo topológicos, complexos, flutuantes, indefinidos, rizomáticos" (PARENTE, 1999, p. 79). O presente trabalho usa elementos de redes complexas como método para atingir os objetivos propostos, tendo como foco a descoberta de relacionamentos entre os termos de consulta do usuário, guiado por métricas de rede.

Quanto a uso de mapas conceituais, que corresponde à última etapa do modelo proposto, Orrantia (2012) em sua pesquisa mostra o quanto eles auxiliam na disseminação da informação. De acordo com Vekiri (2002), há um esforço mental menor para uma compreensão de um texto quando este está acompanhado de diagramas e gráficos, como no caso dos mapas conceituais que possuem uma representação visuo-espacial de um determinado assunto. O’Donnel (1993) conclui em sua pesquisa que há maior eficiência em buscas de informações realizadas pelo usuário quando o conhecimento da base de dados é organizado no formato de mapas de conhecimento em comparação com o formato puramente textual. Hoffman e Beach (2013) citam vários exemplos de sucesso de uso de mapas conceituais em uma variedade de domínios, incluindo a previsão do tempo, oncologia clínica e análise do terreno.

Considerando os mapas conceituais como diagramas, enxergam-se benefícios enquanto forma de apresentar ideias. Ware (2010) enfatiza a importância dos diagramas para concretizar ideias e também como elemento determinante na atividade de apresentar informações e conhecimento. Hook e Börner (2005) argumentam sobre a importância dos mapas para representação do conhecimento em domínios acadêmicos. Green (2012) desenvolveu uma pesquisa com os mapas conceituais para analisar o compromisso dos políticos em suas práticas de desenvolvimento sustentável. Nessa pesquisa, os mapas

conceituais serviram como principal elemento para uma investigação qualitativa que sintetizou e organizou grande quantidade de dados coletados, revelando interconexões entre conceitos, e servindo de base para que políticos e população pudessem aumentar o conhecimento e a compreensão da implementação do desenvolvimento sustentável. Apesar da leitura de mapas conceituais ser bastante intuitiva, existem projetos que estão fazendo uma grande disseminação desse modo de apresentar informações, tal como o projeto World of Science, explicado por Novak e Cañas (2008), onde livros didáticos de ensino de Ciências estão sendo reescritos para que o conteúdo seja trabalhado por intermédio de mapas conceituais prontos e semiprontos a serem completados pelos alunos através de uma investigação conduzida.

Valerio, Leake e Cañas (2012) mostraram que os mapas gerados permitem aos usuários melhorar substancialmente suas habilidades de compreensão de leitura no quesito velocidade em comparação a leitura somente de texto. Nessa mesma linha Lima (2004) argumenta que o "[...] mapa conceitual, com sua característica gráfica, é um instrumento poderoso para se compreender as relações entre os conceitos do conhecimento no todo [...]" ele é uma "[...] ferramenta apropriada para organizar e representar um domínio do conhecimento, auxiliando a externalização das estruturas cognitivas dos autores de hiperdocumentos". De fato, Zhang (2008) observa que sem o auxílio de visualização gráfica, há necessidade de maior abstração de informações e, consequentemente, menor percepção ou compreensão dos dados e informações.

Alguns desafios se destacam na construção de um mapa conceitual. Cañas *et al.* (2004) chama atenção para o grande desafio de criar as ligações entre os conceitos, já que os autores consideram que a seleção ou criação de conceitos é uma tarefa bem mais simples do que o estabelecimento de relações significativas para criar uma estrutura coerente que reflita o entendimento de uma pessoa num domínio. Outro elemento desafiador nos últimos anos é a geração de mapas conceituais de forma automática. Kowata, Cury, Boeres (2010), através de uma análise realizada sobre 15 trabalhos nesse contexto, concluíram que, apesar da geração automática ser bastante desejada, ela tem sido muito mais semiautomática com forte dependência humana.

O modelo proposto mais adiante no presente trabalho utiliza-se de mapas conceituais justificados pelos elementos apresentados nessa seção. Ou seja: (i) a visualização das relações descobertas entre os termos fornecidos pelo usuário no formato de mapa conceitual diminui a sua carga cognitiva e necessidade de abstração ou esforço para interpretar e extrair informações sintetizadas; (ii) maior eficiência na leitura e compreensão das informações

extraídas da base de dados ligados e apresentadas num formato visuo-espacial como mapa conceitual; (iii) os mapas gerados auxiliam a disseminação de informações contidas nas bases de dados ligados; (iv) o foco do modelo na geração das frases de ligação entre os conceitos, apesar de desafiador, contribui com o usuário para o entendimento de relações entre as várias áreas do conhecimento contidas nos termos iniciais da busca; (v) o mapa resultante é gerado de forma totalmente automática poupando tempo do usuário. Além disso, a apresentação dos novos conceitos intermediários aos termos fornecidos pelo usuário no mapa conceitual resultante tem seu entendimento facilitado uma vez que eles vão ao encontro da ideia da Aprendizagem Significativa de Ausubel (1968): os novos conceitos se relacionam com os termos inicialmente fornecidos pelo usuário, que fazem o papel dos subsunçores, permitindo a aproximação da já existente estrutura de conhecimento do usuário. Isto tudo, é claro, se o usuário possuir, mesmo que parcialmente, entendimento sobre os termos fornecidos.

A presente tese contribui para a Ciência da Informação (CI) no fortalecimento de uma de suas premissas básicas: área fortemente interdisciplinar. Isso ocorre pela integração conceitual da recuperação de informação e conhecimento, redes complexas, mapas conceituais, e dados ligados na Web Semântica. O trabalho também contribui com a CI sugerindo um modelo para a recuperação de informação e conhecimento no contexto da web, sinalizado aqui por alguns autores como sendo uma área de gande demanda por melhoria dos métodos empregados. Além disso, por intermédio desse modelo, o trabalho abre discussão, na seção 5.1, sobre uma visão integrada de alguns paradigmas da CI, e interpretações aplicadas da equação fundamental da CI de Brookes.

1.4 Trabalhos correlatos

Investigou-se na literatura trabalhos correlatos com a presente tese. O método usado nessa investigação foi baseado em pesquisa bibliográfica realizada no Portal de Periódico da Capes tendo como base combinações dos seguintes termos de busca: *information retrieval, knowledge retrieval, text retrieval, semantic web, linked data, complex network, network analysis, concept map, relationship concept, concept map generation, information visualization, knowledge visualization*, e seus correspondentes na língua portuguesa.

Foram identificados e selecionados os trabalhos que atendiam a um mínimo de quatro critérios em comum com o presente trabalho. Ao todo foram estabelecidos 13 critérios comparativos de acordo com características julgadas relevantes sobre o presente trabalho. Os

critérios, bem como a análise comparativa dos trabalhos selecionados com o presente trabalho são apresentados no capítulo de discussão, na subseção 5.8.

Os próximos parágrafos, organizados cronologicamente, descrevem brevemente os trabalhos selecionados para a análise. Termos e assuntos específicos de determinadas áreas, usados nas descrições, podem ser consultados no referencial teórico, no capítulo 2.

O software *Concept Suggester* (CAÑAS *et al.*, 2004) sugere conceitos para um mapa conceitual que está sendo construído. As sugestões são buscadas na web e extraídas a partir de análise feita no mapa em construção. O usuário decide entre incluí-las ou não no mapa. A abordagem para extração dos conceitos sugeridos é baseada na busca de cada conceito do mapa conceitual na web e, para cada documento achado com esse conceito, a seleção de todas as palavras vizinhas desse conceito. A seleção final da lista de termos a serem sugeridos é um ranqueamento das palavras mais frequentes em toda as seleções realizadas. Como maior benefício, esse sistema permite que o usuário tenha maior foco na criação das frases de ligação do que na busca por conceitos, já que essa última tarefa requer um esforço cognitivo mais simples. Os autores detectaram que o algoritmo proposto conseguiu ser mais efetivo em estágios iniciais da construção de mapas conceituais. Quando o mapa tinha uma quantidade maior de conceitos o algoritmo não conseguia selecionar bons conceitos para o usuário.

Lima (2005) criou um protótipo digital, denominado Modelo Hipertextual para Organização de Documentos (MHTX), que consiste em um mapa semântico, modelado como mapa conceitual, e em um sumário expandido. Esse sistema auxilia a organização e representação do conhecimento humano em hipertextos e usa mapas conceituais para "[...] possibilitar ao usuário uma visão geral da estrutura semântica do texto escolhido através de sua representação gráfica, facilitando a navegação semântica em contexto, através de seções e subseções, digitalizadas, constantes da base de dados hipertextual" (p. 6). A autora concluiu que o mapa conceitual foi uma alternativa viável para o problema de desorientação do usuário. O sistema oferece uma navegação orientada que explicita o conteúdo semântico e suas conexões.

Thammasut e Sornil (2006), em sua pesquisa, desenvolveram um sistema de recuperação de informação onde os documentos são ranqueados de acordo com os relacionamentos existentes entre os documentos, tópicos e sub-tópicos e os termos da consulta. O usuário especifica a sua consulta de forma intuitiva por intermédio de um grafo. Depois, esse grafo é consolidado com o grafo formado pelos documentos. O ranking dos documentos é calculado a partir desse grafo consolidado e usando um algoritmo denominado de Random Walks with Restart.

Truong *et al.* (2008) propuseram um método para RI baseado na medição de similaridades entre vértices de dois grafos. Documentos, consultas e termos de indexação são vértices de um grafo bipartido onde arestas ligam documentos e consultas com termos de indexação. Os documentos recuperados são nós considerados similares aos nós da consulta. Os autores ainda observaram que o método proposto possui alta complexidade computacional para bases de conhecimento muito grandes.

Graudina e Grundspenkis (2008) observaram que tanto mapas conceituais como ontologias representam algum domínio, pois ambos possuem classes ou conceitos e relações entre eles, porém ontologias possuem atributos para classes e, assim, são mais expressivas. Contudo os autores fizeram uma correspondência entre os vários elementos que compõe uma ontologia com os elementos dos mapas conceituais para aumentar a expressividade desses últimos. Dessa forma, eles desenvolveram algoritmos para criação de mapas conceituais de forma automática a partir de ontologias escritas em OWL.

Heim, Ertl e Ziegler (2010) propõem o projeto gFacet que possui uma ferramenta para realizar consulta semântica através de uma abordagem que explora dados RDF da DBpedia pela combinação de visualizações baseadas em gráficos com técnicas de filtragem facetada. A filtragem facetada é representada como um nó numa visualização gráfica e pode ser interativamente adicionada ou removida pelo usuário, a fim de produzir interfaces de pesquisa individuais. A ferramenta usa consultas SPARQL para acessar bancos de dados RDF.

Lohmann *et al.* (2010) desenvolveram a ferramenta RelFinder para descoberta interativa de relacionamentos de entidades num contexto sobre dados ligados com RDF. A ferramenta extrai e visualiza relações entre os objetos fornecidos pelo usuário e os apresenta, juntamente com suas relações, num formato visual de grafo e ainda permite manipulações interativas com os objetos descobertos. Os relacionamentos são encontrados com base em um algoritmo proposto por Lehmann, Schüppel e Auer (2007), e por intermédio de sucessivas consultas SPARQL (HEIM *et al.*, 2009). A ferramenta RelFinder também permite várias configurações no grafo em um nível detalhado ou global, como a determinação da quantidade de nós e relações, os tipos de relacionamentos que devem ser apresentados além de resgatar informações sobre um determinado nó escolhido.

Guéret *et al.* (2012) desenvolveram o LINK-QA, uma estrutura extensível para a realização de avaliação de qualidade na Web de Dados. Eles descreveram cinco métricas para determinação da qualidade de *Linked Data*. Essas métricas foram analisadas usando conjuntos conhecidos de ligações boas e ruins criadas com uma ferramenta de mapeamento apropriada. Porém, as métricas de rede demonstraram que foram parcialmente eficazes na detecção da

qualidade dessas ligações. Os autores concluíram que as medidas de rede não foram suficientes para a avaliação proposta e precisam ser mais estudadas.

Valerio, Leake e Cañas (2012) propõe um algoritmo para geração automática de mapas conceituais com o intuito de resumir o conteúdo de documentos e, dessa forma, facilitar a tarefa humana de compreensão deles. Eles usaram ferramentas de processamento de linguagem natural para extrair informações dos documentos, gerar árvores de análise para a construção de uma estrutura hierárquica que é convertida em um mapa conceitual. Os autores conseguiram mostrar que os mapas gerados permitiram a melhoria substancial das habilidades dos usuários de compreensão e leitura no quesito velocidade.

Paulheim (2013) apresenta um método para mineração de dados em *Linked Data* usando uma estratégia de pré-processamento sobre os dados ligados. Dessa forma, é possível a utilização de muitos algoritmos e ferramentas já existentes para a mineração de dados convencional. A ferramenta proposta aceita entrada de texto e usa SPARQL para acesso a base dados ligados adotada: DBpedia. O autor agregou conhecimento nas informações recuperadas para aumentar a semântica e, assim, obteve melhores resultados.

McLinden (2013) realizou um trabalho conjunto entre mapas conceituais e análise de redes sociais e concluiu que, apesar dos objetivos dos mapas conceituais serem diferentes das redes de informação, as estruturas de dados subjacentes são semelhantes, pois em ambos os casos existe uma rede de relações entre elementos de dados. A entrada é realizada pelo formato de redes sociais. O autor usa métricas de rede para selecionar e agrupar informações das redes sociais, segundo alguns critérios, com o objetivo de auxiliar a construção dos mapas conceituais resultantes.

Cury, Perin, Santos Junior (2014) criaram a plataforma orientada a serviço CMPAAS que usa a arquitetura conhecida por SOA[4] e oferece uma base para suportar a interoperabilidade em funções criadas para diversas manipulações sobre mapas conceituais. Essas funcionalidades podem ser criadas de forma independente por qualquer um e de qualquer lugar do mundo. Alguns serviços sobre mapas conceituais já funcionam como protótipos, tais como, edição, armazenamento em repositórios, comparação, mesclagem, geração automática etc. Outros serviços caminham na direção de proporcionar análise, processamento estatístico e avaliação em mapas conceituais.

[4] *Service Oriented Arquitecture* ou arquitetura orientada a serviços é um estilo de arquitetura de software cujo princípio fundamental é que as aplicações devem ser disponibilizadas na forma de serviços. Ela permite a criação de serviços de negócio que, apesar do desenvolvimento independente, são interoperáveis e com facilidade na reutilização e compartilhamento entre aplicações e empresas.

Usbeck (2014) propôs um sistema de informação como um ambiente e máquina de busca. A principal questão de seu trabalho, que ainda está em andamento, é como uma máquina de busca pode se beneficiar do paradigma *Linked Data*? Usando algoritmos altamente escaláveis, o sistema combina as vantagens de métodos de recuperação de informação e tecnologias *Linked Data* pra superar o problema de grande número de informações disponíveis. Num escopo reduzido, ele também extrai informações de páginas web não estruturadas para formação dos dados ligados que, posteriormente, poderão ser usados na busca. Sobre o ranqueamento das informações recuperadas, o autor ainda está desenvolvendo e deve utilizar aprendizado de máquina bem como vários algoritmos de classificação conhecidos.

2 REFERENCIAL TEÓRICO

Esse capítulo apresenta os elementos teóricos que fundamentam e são referência para o presente trabalho. Em função disso, e também com o intuito de oferecer uma ordem adequada aos tópicos, foi feita uma organização que contemplasse de forma sintética e didática todos assuntos do trabalho, cobertos por várias áreas do conhecimento.

A primeira seção, a mais abrangente, aborda a Ciência da Informação (CI), com foco maior em elementos que fundamentam o modelo proposto, como os paradigmas da CI e a equação fundamental da CI. A segunda seção aborda a formação de conceito e suas relações; a caracterização, criação e aplicações dos mapas conceituais; e a aprendizagem significativa. A terceira seção trata da Web Semântica com foco maior nos dados abertos ligados. A quarta seção aborda a recuperação de informação (RI), que apesar de ser uma área de estudo da CI, foi aqui destacada como uma seção a parte devido a sua importância no trabalho; além disso, aborda áreas de interesse da RI, tais como a recuperação de conhecimento e a visualização de informação e conhecimento. A quinta seção apresenta a Ciência das Redes, porém, trata apenas dos elementos diretamente relacionados ao desenvolvimento da tese, tendo em vista a vasta literatura e a intensa expansão deste campo de pesquisa. Ao final de cada seção são tecidas considerações para retomar alguns pontos discutidos de forma sintética e integrada com os vários assuntos tratados. Finalmente, a sexta e última seção faz uma abordagem de tópicos fundamentais de todo o referencial teórico revelando importantes pontos de integração entre as várias áreas de conhecimento apresentadas no referencial teórico.

Dentre os autores e assuntos fundamentais nesse referencial teórico, destacam-se os seguintes:

(i) Paradigmas da CI (CAPURRO, 2003);
(ii) Equação fundamental da CI (BROOKES, 1980c);
(iii) Teoria do conceito (DAHLBERG, 1978);
(iv) Proposição (RUSSELL, 1919);
(v) Teoria da aprendizagem significativa (AUSUBEL, 1968);
(vi) Mapas conceituais (NOVAK, 1977);
(vii) Web Semântica (BERNERS-LEE *et al.*, 2001);
(viii) Dados ligados (BERNERS-LEE, 2006);
(ix) Recuperação de informação (BAEZA-YATES; RIBEIRO-NETO, 2011);
(x) Recuperação de conhecimento (YAO *et al.*, 2007);
(xi) Visualização de conhecimento (BURKHARD, 2005);

(xii) Avaliação da informação recuperada (CLEVERDON, 1962);
(xiii) Teoria dos grafos por Leonard Euler em 1736;
(xiv) Fenômenos em redes complexas (BARABÁSI, 2002); e
(xv) Métricas de rede (NEWMAN, 2003, 2010).

Outras contribuições serão citadas ao longo desse referencial teórico.

2.1 Ciência da Informação

Nos anos 60 a Ciência da Informação (CI) se consolidou teórica e institucionalmente, primeiramente nos Estados Unidos, União Soviética e Inglaterra e, na década seguinte em vários outros países (ARAÚJO, 2014). Diferentemente de outras ciências mais antigas e consolidadas, ela já nasceu como um campo de conhecimento interdisciplinar tendo essa característica muito forte e destacada por diversos autores, entre eles, Saracevic (1995), Le Coadic (1996), Robredo (2003b), Pinheiro (2007), Bicalho (2010), Beluzzo *et al.* (2011), e Araujo (2014).

Shera e Cleveland (apud FONSECA, 2005, p. 19) resgatam uma das primeiras definições de CI, elaborada em 1962 numa conferência realizada no Georgia Institute of Technology: "Ciência que investiga as propriedades e o comportamento da informação, as forças que governam os fluxo de informações e os meios de processar a informação para otimizar o seu acesso e uso". Essa definição é avalizada por diversos pesquisadores citados em Beluzzo *et al.* (2011). De forma ainda mais concisa, Robredo (2003b, p. 105), a define como "[...] é o estudo, com critérios, princípios e métodos científicos, da informação". Le Coadic (1996) afirma que a CI tem como objetivo "[...] o estudo das propriedades gerais da informação (natureza, gênese, efeitos), e a análise de seus processos de construção, comunicação e uso". Porém, Capurro e Hjørland (2003) citam Schrader, que estudou aproximadamente 700 definições de CI entre 1900 e 1981 e disse que a literatura de CI é caracterizada pelo caos conceitual. Essa diversidade pode ser um ponto positivo no sentido dessa ciência ter condições de atender demandas em várias áreas do conhecimento.

Essa seção explora características e abrangência da CI, os paradigmas da CI, os termos informação e conhecimento e seus desdobramentos e a equação fundamental da CI de Brookes.

2.1.1 Características e abrangência

A CI tem uma forte dimensão social e humana (SARACEVIC, 1995) sendo baseada na noção das necessidades de informação de pessoas envolvidas em trabalhos sociais, se relacionando com métodos de organização de processos de comunicação de forma a atender essas necessidades informacionais (VICKERY; VICKERY, 1987; WERSIG; NEVELING, 1975). Ela também é baseada no processo de "[...] comunicação e informação que se desenvolve em diferentes territórios: científicos, tecnológicos, educacionais, sociais, artísticos e culturais, portanto, múltiplos contextos e condições experimentais" (PINHEIRO, 2007, p. 98). Essa diversidade também é destacada por Araujo (2014) que completa dizendo que os modelos de compreensão distintos, campos de estudo diversos e variados objetos empíricos da CI têm evidenciado a inexistência de um corpo teórico unificado e acabado, que, ao invés de indicar um sintoma de imaturidade ou fragilidade da área, tem sido entendido como um aspecto intelectualmente estimulante para exercer a criatividade na formulação de conceitos e compreensão de novos fenômenos. Araujo ainda destaca que essas características colocam a CI num patamar de amplo diálogo com as mais distintas áreas disciplinares. Nessa mesma linha, Saracevic (1995) sinaliza três características na CI que são os motivos de sua própria evolução e existência:

(i) É interdisciplinar;

(ii) Está ligada à tecnologia da informação; e

(iii) É uma ciência participante ativa na evolução da sociedade da informação.

A partir dessas caracterizações da CI percebe-se que ela corresponde a um universo muito amplo de estudo e atuação, e com possibilidades de contribuição em várias áreas do conhecimento. Almeida Junior (2009, p. 46) destaca a diferença da CI das outras áreas: todas lidam com a informação, mas "[...] o objeto da Ciência da Informação não é a informação em si, mas a mediação dela". Beluzzo *et al.* (2011) consideram que a CI se situa no contexto das ciências pós-modernas, interdisciplinares, sendo um dos principais meios de acesso a uma compreensão do social e do cultural.

Floridi (2002) apresenta a CI como estreitamente relacionada à Epistemologia Social e a Filosofia da Informação, sendo essa última entendida como uma filosofia fundamental da análise de informações e *design*. O autor também argumenta que a Epistemologia Social não consegue oferecer uma fundamentação para a CI, enquanto, por outro lado, a CI é uma Filosofia da Informação Aplicada. Alguns autores destacaram a ampla atuação da CI. Por

exemplo, Vickery e Vickery (1987) argumentam que ela auxilia a compreensão em seis grandes áreas:

(i) O comportamento das pessoas como geradores, fontes, destinatários e usuários da informação, e como agentes de canal;

(ii) O estudo quantitativo da população de mensagens de seu tamanho, taxa de crescimento, distribuição, padrões de produção, e utilização;

(iii) A organização semântica das mensagens e dos canais que facilita sua identificação por fontes e receptores;

(iv) Problemas associados com as funções de armazenamento de informação, análise e recuperação;

(v) A organização geral dos sistemas de informação e seu desempenho na transferência; e

(vi) O contexto social de transferência de informação, em particular na economia e política.

Araujo (2014) destaca três grandes propostas levantadas para a CI: (i) ela é interdisciplinar sendo essa uma característica natural da área; (ii) ela é pós-moderna no sentido do seu principal objeto de estudo, a informação, não ser novo, mas já amplamente conhecido pelas outras ciências; e (iii) ela é uma ciência humana e social, sendo uma das consequências disso, é que a informação não existe independente dos sujeitos que se relacionam com ela. Contudo, Capurro (2003) alerta que para tratar de um campo mais específico da CI deve-se transportar a sua definição para um nível mais abstrato fazendo uma reflexão epistemológica.

2.1.2 Paradigmas da Ciência da Informação

Capurro (2003), em seu artigo, faz um levantamento de alguns paradigmas epistemológicos que influenciaram a Ciência da Informação, tais como hermenêutica, racionalismo crítico, semiótica, construtivismo, cibernética de segunda ordem e teoria de sistemas. O autor destaca também duas raízes para a Ciência da Informação, uma é a Biblioteconomia clássica ou a Ciência das Mensagens ligada a todos os aspectos sociais e culturais do mundo humano e, a outra, é referente ao “[...] impacto da computação nos processos de produção, coleta, organização, interpretação, armazenagem, recuperação, disseminação, transformação e uso da informação, e em especial da informação científica

registrada em documentos impressos" (p. 7) . Porém, de forma mais sistematizada e com foco maior, três paradigmas principais para a Ciência da Informação foram identificados por Capurro: físico, cognitivo e social. Nascimento (2006), no Quadro 1, sintetiza esses três paradigmas caracterizando-os quanto a abordagem, ao tipo do processo e o foco (o olhar).

Quadro 1 – Principais paradigmas da Ciência da Informação

Paradigmas	**Abordagem**	**Processos**	**O olhar**
Cognitivo	Indivíduo	Psicológicos	Organização e tratamento da informação
Físico	Sistema	Tecnológicos	
Social	Domínio	Sociais e culturais	Informação construída

Fonte: (NASCIMENTO, 2006)

Compilando Capurro (2003), Nascimento (2006) e Araujo (2014), os três paradigmas podem ser resumidos:

- **Paradigma Físico**. Tem abordagem focada nos sistemas de informação por intermédio de processos tecnológicos. Baseado na recuperação de informação onde a informação é algo, um objeto físico, onde um emissor transmite a um receptor. Ele tem suas raízes e seu sentido em atividades clássicas dos bibliotecários e documentalistas. Está relacionado à Teoria da Informação de Claude Shannon e Warren Weaver, de 1949.
- **Paradigma Cognitivo**. Tem abordagem no indivíduo por intermédio de processos psicológicos. A informação se relaciona às estruturas de conhecimento do usuário. Porém, parte da premissa de que a busca de informação tem sua origem na necessidade do usuário onde o conhecimento que está ao seu alcance não é suficiente para satisfazer tal necessidade. Proposto inicialmente por Bertram C. Brookes em 1977, que foi diretamente influenciado pela epistemologia de Karl Popper com os seus três mundos[5].
- **Paradigma Social**. O usuário é considerado de uma forma inserida em contextos sociais onde a informação é dependente do domínio a qual ela pertence. Tendo foco na informação construída, um conhecimento só faz sentido se houver um pressuposto conhecido e compartilhado com outros. Esse paradigma dá a "[...] possibilidade de

[5] Três mundos de Karl Popper: mundo das coisas físicas, mundo do conhecimento subjetivo e mundo do conhecimento objetivo. A seção 2.1.4, sobre a equação fundamental da CI, aborda com mais detalhes.

> provocar a produção de outro conhecimento, não linear, mas circular, que valorize as inter-relações culturais, ambientais, sociais, econômicas e políticas construídas para enfrentar de forma mais coerente e atuante os desafios atuais da sociedade" (NASCIMENTO, 2006, p. 33). Tem como principais expoentes a filosofia de Ludwig Wittgenstein em 1958, a Teoria do Discurso de Michel Foucault, de 1994, e os estudos de Birger Hjørland e Hanne Albrechtsen a partir de 1995.

Capurro (2003) ainda alerta que a construção desses paradigmas não decorreu de um processo linear histórico. Cada um dos paradigmas veio sendo construído ao longo da história e de forma concomitante. Araujo (2014) destaca várias outras classificações, modelos e paradigmas para a CI, entre eles, os modelos positivista, cognitivo e sociológico de Fernandéz Molina e Moya-Anegón e a sistematização de Ørom (2000 apud ARAÚJO, CARLOS ALBERTO ÁVILA, 2014, p. 20) através dos aspectos físico, semântico e pragmático. Ambos modelos e aspectos tem uma proximidade muito grande aos paradigmas apresentados por Capurro. Araujo ainda propõe uma sistematização dos três paradigmas como: (i) objetivo, onde o foco está voltado para a "[...] construção de modelos e sistemas que garantam um transporte mais rápido, mais barato e mais eficiente das mensagens ou sinais que são trocados entre diferentes sujeitos" (p. 23); sendo uma abordagem que ainda é atual e tem como exemplos os motores de busca na internet; (ii) subjetiva, que se aproxima do modelo cognitivo buscando uma dimensão informacional no processo; (iii) intersubjetivo, numa "[...] perspectiva contemporânea pragmatista, insere-se no contexto sociocultural e a dimensão interacional dos sujeitos no escopo do objeto de estudo do campo" (p. 24).

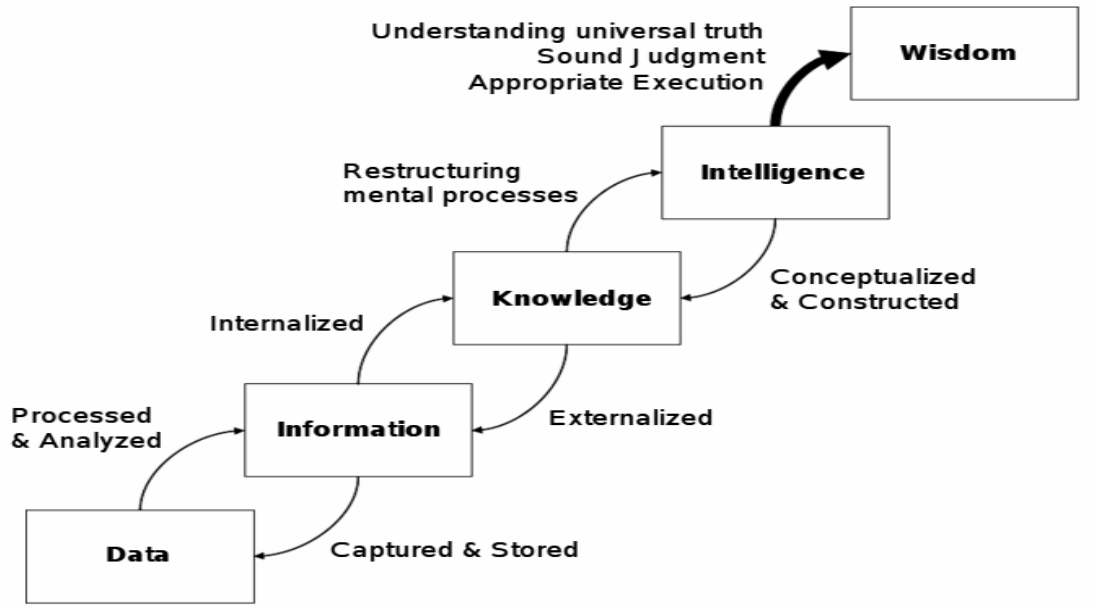
De modo geral, os paradigmas da CI, tal como adotados por diversos pesquisadores, giram em torno dos três apontados por Capurro e destacados nessa subseção, sendo uma tendência atual da CI seguir o Paradigma Social (HJØRLAND, 2010). Além disso, pode-se observar que essas abordagens são mais complementares do que excludentes e é o encontro delas que pode definir de forma consistente a CI (ARAÚJO, 2014).

2.1.3 Informação e conhecimento

Dado, informação e conhecimento são três conceitos fundamentais e inter-relacionados entre si no contexto da CI (ZINS, 2007), sendo informação um conceito interdisciplinar onde quase toda disciplina científica a define de forma contextualizada a fenômenos específicos (CAPURRO; HJØRLAND, 2003). No senso comum eles são

representados por uma tríade piramidal que muitas vezes é até estendida para: sinal-símbolo-dado-informação-conhecimento-inteligência-sabedoria, apresentando o elemento mais primitivo em sua base - o sinal - até o elemento mais elaborado, em seu topo - a sabedoria. O relacionamento entre os elementos adjacentes dessa pirâmide transmite a ideia de composição, ou seja, um símbolo é formado a partir de um conjunto organizado de sinais, uma informação é composta de um conjunto organizado de dados, e assim por diante. Porém, essa hierarquização não pode ser simplesmente generalizada para todos os elementos dessa pirâmide. De forma sintética, Liew (2013) estabelece relações circulares, na Figura 1, onde dado processado e analisado é uma informação; já a informação, quando internalizada na mente, é conhecimento; o conhecimento, quando reestruturado num processo mental, é considerado inteligência e, finalmente, a inteligência, entendida como verdade universal, julgada de forma segura e executada adequadamente, é sabedoria. Por outro lado, inteligência, quando conceitualizada e construída, gera conhecimento; este, quando externalizado, gera informação; esta, quando capturada e armazenada, gera dados.

Figura 1 – Relacionamentos entre data, informação, conhecimento, inteligência e sabedoria



Fonte: Liew (2013)

Porém, apenas descrever as relações existentes entre eles não é o mesmo que defini-los (LIEW, 2007, 2013). Existem muitas controvérsias, por exemplo, Zins (2007) alerta que se informação fosse simplesmente um conjunto de dados organizado então isso sugeriria que a CI abordasse somente dados e informações, mas não conhecimento, já que informação é um

conjunto de dados e conhecimento pertenceria a um nível superior. Kuhlen (1996, apud. CAPURRO, 2003) vê o relacionamento entre informação e conhecimento oposta do senso comum, ou seja, informação é conhecimento em ação, que, de certa forma, é compatível com a ideia de Liew (2013), quando afirma que informação é o conhecimento externalizado. Enquanto Bretx (1971, apud CASADO, 1994) afirma que a transformação de informação em conhecimento necessita de estruturas preexistentes na memória do indivíduo que são capazes de reter informações para formar conhecimento. Nessa mesma linha, a teoria da Aprendizagem Significativa, formulada por Ausubel (1968), estabelece que uma nova informação se relaciona com a estrutura do conhecimento do indivíduo para formar novos conhecimentos.

Shannon (1948), ao propor a Teoria da Informação, definiu a informação como sendo algo numérico, lógico, capaz de ser mensurado, pois dessa forma seria possível quantificar ruídos numa transmissão e medir a qualidade desse processo. Porém, ele deixou de fora dessa definição preocupações relativas à semântica ou a intensão do emissor da informação (CAPURRO; HJØRLAND, 2003; LOGAN, 2014). Por outro lado, Donald Mackay (1951 apud LOGAN, 2014) defendeu a ideia de que informação não poderia ser dissociada de seu significado e só poderia ser definida em correlação a um contexto mais amplo. Contudo, Logan desenvolve em sua obra a ideia de que a informação é relativista, ou seja, a sua definição é dependente do contexto ao qual se aplica. Além disso, ele relacionou informação à propagação da organização e, dessa forma, concluiu que ela é mais um processo ou um verbo do que um substantivo.

Enquanto Le Coadic (1996) dá um sentido semântico à informação afirmando que ela é um "[...] conhecimento inscrito (gravado) sob a forma escrita (impressa ou numérica), oral ou audiovisual", Buckland (1991), numa visão pragmática, identificou três usos: (i) informação como um processo, relaciona-se ao ato de informar, então se alguém é informado de algo, o seu conhecimento modifica-se; (ii) informação como um conhecimento, é o conhecimento comunicado referente a algum fato, assunto ou evento; (iii) informação como coisa, usado para atribuir objetos tais como dados e documentos. Saracevic (1999) identifica três sentidos para a informação: (i) sentido restrito, onde a informação é considerada em termos de sinais ou mensagens que envolvam pouco ou nenhum processamento cognitivo; (ii) sentido amplo, onde a informação é tratada como algo que envolve diretamente o processamento cognitivo e a compreensão, sendo resultado da interação de duas estruturas cognitivas, e sendo um bem intangível que depende da conceituação e da compreensão de um ser humano; (iii) sentido ainda mais amplo, onde a

informação é tratada em um contexto, não envolvendo apenas as mensagens (sentido estrito) que são cognitivamente processados (sentido amplo), mas, também o contexto da situação, tarefa ou problema em questão.

Zins (2007) documentou e analisou 130 definições para dados, informação e conhecimento realizadas por 45 acadêmicos na área da Ciência da Informação. O autor destacou que essas definições se enquadram em duas categorias principais: o domínio subjetivo (internalizado em um indivíduo) e o domínio universal (coletivamente externalizados de modo objetivo). Após análise de todas as contribuições, Zins identificou a existência de cinco modelos diferentes sobre as possibilidades de enquadramento dos conceitos para dado, informação e conhecimento nos dois domínios. Ele ainda observou que, dado foi classificado predominantemente pertencente ao domínio universal, enquanto conhecimento foi caracterizado predominantemente como pertencente ao domínio subjetivo. Contudo, informação não obedeceu a um padrão, alternando-se entre os dois domínios ou pertencendo aos dois.

Silva e Gomes (2015) também analisaram manifestações do conceito de informação na CI em publicações de 1971 a 2008, 21 ao todo, e observaram que elas não representam a "[...] quantidade totalizante das definições apresentadas na CI, mas representam uma expressiva qualidade conceitual de cunho conteudístico e semântico na CI [...]" e, dessa forma, elas conseguem abranger "[...] fundamentos científicos (teor lógico-epistemológico), humanos (intercorrência sociais) e técnico-pragmáticos (empíricos)" (p. 148). Após várias considerações, interpretações e análises as autoras propuseram o seguinte conceito semanticamente geral de informação:

> A informação é uma produção fenomenicamente social que tem por finalidade dinamizar a inter-comunicação humana e promover exposições e descobertas para construção do conhecimento através de interações entre sujeito/autor e sujeito/usuário por meio de dados (plano físico e histórico-social dos sujeitos da informação), mensagens (no plano abstrativo) e atividades documentais (plano material), que favorecem predicativos hermenêuticos aos sujeitos da informação e resultam na apreensão e apropriação pelo sujeito/usuário efetivando um caráter de compreensão (p. 150).

A definição proposta pelas autoras parece abarcar as várias necessidades do uso da informação em vários contextos, satisfazendo os planos físico, histórico-social, abstrativo e material. No entanto, Capurro e Hjorland (CAPURRO; HJØRLAND, 2003) destacam que na CI é importante distinguir informação entre dois polos, como um objeto ou coisa, e como um conceito subjetivo e dependente da interpretação de um agente cognitivo, ou seja, o domínio e o contexto onde se usa a informação são determinantes para a sua conceituação. Os autores exemplificam o caso de uma pedra encontrada num campo. Ela representa um tipo de

informação diferente para um arqueólogo e para um geólogo, por exemplo. Dessa forma, Brookes (1980c) tem o conceito de informação no contexto de sua equação, que será abordada em detalhes na seção 2.1.4, como sendo um elemento que provoca transformações nas estruturas cognitivas do sujeito, e quando se relaciona à estrutura cognitiva existente desse sujeito, produz novo conhecimento. Brookes ainda observou o conhecimento como sendo uma estrutura de conceitos ligados por suas relações e informação como uma parte pequena dessa estrutura. Essa estrutura de conhecimento, segundo Brookes, pode ser objetiva ou subjetiva. Sobre a estrutura de conceitos ligados, Bouding (1956 apud. LE COADIC, 1996, p. 9) afirma na mesma linha de Brookes, que nosso estado de conhecimento sobre determinado assunto, em determinado momento, é representado por uma estrutura de conceitos ligados por suas relações.

2.1.4 A equação fundamental da Ciência da Informação de Brookes

Bertram C. Brookes escreveu uma série de quatro artigos (BROOKES, 1980a, 1980b, 1980c, 1981) para discutir vários fundamentos da CI. Ele propôs uma equação fundamental para representar a relação entre a informação e o conhecimento (BROOKES, 1980c) que, mais adiante ficou conhecida como a 'Equação Fundamental da Ciência da Informação', equação 1.

$$K[S] + \Delta I = K[S+ \Delta S] \quad (1)$$

A equação 1 teve sua importância confirmada por vários autores como Neill (1982), Le Coadic (1996), Todd (1999), Robredo (2003a), Araújo (2003), Nascimento (2006), Batista Costa e Alvares (2007), Pereira (2008), Bawden (2011), Moraes (2013), Pontes Junior, Carvalho e Azevedo (2013). Mesmo que alguns desses autores relatassem algumas discordâncias quanto a ideias associadas à equação, todos reconheceram a sua relevância. Em um levantamento realizado por Pereira (2008), de 1980 a 2008, 106 autores apresentaram trabalhos que citaram os quatro artigos de Brookes em suas referências.

Brookes teve como base para seu estudo os três mundos de Karl Popper[6]: (i) Mundo 1: é o conjunto de todas as coisas físicas; (ii) Mundo 2: é conhecimento humano subjetivo ou estados mentais; e (iii) Mundo 3: é o conhecimento objetivo, ou seja, produtos da mente humana porém armazenados em artefatos. Brookes lembrou que, embora esses três mundos

[6] Karl Raimund Popper (1902-1994): filósofo da ciência, austríaco e naturalizado britânico.

sejam independentes, eles possuem interação entre si, sendo que ele defendeu a ideia de que o domínio da CI é o estudo das interações entre os mundos 2 e 3. Brookes afirmou que a sua equação se aplica a ambas estruturas de conhecimentos subjetivo e objetivo, ou seja, o mundo 2 e o mundo 3 de Popper. Para Brookes, as ideias de Popper tinham grande relevância para a CI (NEILL, 1982).

Brookes formulou a equação 1 seguindo uma linha de estudo que considerou mais fortemente o paradigma cognitivo da CI, tratado na subseção 2.1.2. Ele se baseou na ideia de que a informação provoca transformações nas estruturas cognitivas do indivíduo. Essas estruturas podem ser subjetivas e objetivas, e são formadas por conceitos que estão ligados entre si e com as relações que o indivíduo possui, isto é, a sua imagem de mundo. Assim, K[S] é a estrutura cognitiva do sujeito; ΔI é uma nova informação recebida pelo sujeito que, relacionando-se com a sua estrutura cognitiva atual K[S], provoca alterações representadas por ΔS. A parcela K[S+ ΔS] representa a nova estrutura cognitiva do sujeito após relacionamento com a nova informação ΔI e em função do seu novo estado S + ΔS.

Brookes também observou que a parcela ΔI poderia ser definida como um pequeno pedaço de conhecimento ΔK, como mostra a equação 2.

$$K[S] + \Delta K = K[S+ \Delta S] \quad (2)$$

Porém, ele esclareceu que ΔI pode ter diferentes efeitos sobre diferentes estruturas de conhecimento e, portanto, poderia sem prejuízo permanecer ΔI. Além disso, o autor também deixa claro que a equação não diz que o conhecimento é simplesmente aumentado com a chegada da nova informação, mas, a absorção da nova informação em contato com a estrutura de conhecimento do sujeito causa uma modificação nas relações conceituais já existentes e, portanto, não deve ser admitida como um simples incremento de informação, sendo que a percepção desta é dependente da observação sensorial do sujeito.

Em função disso, alguns autores, tais como Le Coadic (1996), Robredo (2003a), Araújo (2003) e Pereira (2008) passaram a escrever a equação de forma a representar essa observação de Brookes, destacando a transformação de ΔI por ΔK, como mostra a equação 3.

$$K[S] + \Delta K = K[S+ \Delta S] \quad (3)$$

$$\uparrow$$

$$\Delta I$$

Brookes, propositadamente, expressou a equação num formato pseudo-matemático com o intuito de mostrar a relação entre informação e conhecimento de forma compacta. Ele ainda observou que, para os matemáticos, os termos e símbolos de sua equação são indefinidos. Por exemplo, o sinal '=', usado na equação, é mais indicado para representar um equilíbrio do que a igualdade propriamente dita, tal como conhecida na Matemática. Outro exemplo é o sinal de '+' que não representa uma simples adição, mas poderia até mesmo representar uma "[...] disciplina inteira" (NEILL, 1982). Ele ainda destacou que a equação também serve para enfatizar o quão pouco que se sabe sobre as formas nas quais o conhecimento do sujeito cresce.

Brookes (1981) concluiu que a informação oferecida a terceiros deve ser apresentada dentro de uma estrutura cognitiva relevante, pois essa é a única maneira pela qual os usuários podem efetivamente recebê-la. A análise da equação fundamental identifica uma série de pressupostos fundamentais e potenciais hipóteses que são oportunidades para a investigação sistemática detalhada (TODD, 1999). Segundo Pereira (2008) os estudos de Brookes sobre a equação fundamental da CI são importantes e "[...] ainda hoje são utilizados por diversos de seus pares, nas diversas linhas de pesquisa dentro da área da CI, ou como referência principal para o desenvolvimento de novas propostas, ou como base para uma contra-opinião (crítica) e a exposição de uma nova ideia na área" (p. 26). Pereira ainda completa afirmando que "[...] de qualquer forma, a utilização de seus trabalhos como referência é presente e inquestionável" (p. 26).

Araujo (2003) sugeriu uma ampliação da compreensão sobre a relação entre informação e sociedade, usando a equação de Brookes como ponto de partida, pela proposição de uma equação que pudesse explicar o impacto informacional, isto é, conhecer o nível de transformação que ocorre nos sujeitos sociais e nas formações sociais. A equação 4 representa essa ideia, onde: IF - impacto informacional, Ni – as necessidades informacionais, Cs - contextos sociais vivenciados pelos sujeitos, Int – intencionalidade explicita ou não da informação disseminada e/ou utilizada.

$$IF = Ni + Cs \times Int \qquad (4)$$

A equação 4, do impacto informacional, consegue reunir, num só momento, os elementos caracterizadores da transformação/mudança mentais e/ou sociais, segundo a autora: necessidades informacionais dos sujeitos; contextos socioculturais dos sujeitos;

intencionalidade (explicita ou não) da informação disseminada e/ou utilizada. Araujo ainda considera que os dois primeiros elementos têm sido analisados por vários estudos da área, porém, o terceiro elemento surge como algo mais subjetivo e que ainda não tem sido considerado, em termos teóricos conceituais na área da CI.

Por outro lado há quem critique a equação de Brookes e até mesmo a ele próprio quando, por exemplo, Pereira (2008) afirma que a equação serviu para mostrar o quão pouco Brookes sabia, na época, sobre as maneiras em que o conhecimento das pessoas cresce e se desenvolve. Capurro e Hjørland (2003) criticam a definição de informação dada por Brookes através de sua equação, argumentando que é uma definição que "[...] parece-nos servir apenas a essa função persuasiva" (p. 154). Além disso, não é consenso que a equação proposta por Brookes seja considerada a equação fundamental da CI. Por exemplo, Robredo (2003a) disse que o "[...] esquema sugerido por Brookes [...] denominado por ele, um tanto, digamos presunçosamente, equação fundamental da ciência da informação [...]" (p. 14). Apesar disso, Robredo destaca que Brookes acertou ao "[...] reunir a informação (externa ao sujeito detentor de um certo conhecimento), a comunicação (que traz essa informação até o sujeito) e o conhecimento (que se enriquece com a incorporação da informação adicionada)" (p. 16). Porém, o autor sinaliza que falta considerar o processo global de enriquecimento do conhecimento existente por intermédio da "[...] nova informação recebida, de algum componente de natureza psicossomática, tal como o processo específico de análise dessa informação, que permite, entre outras coisas, apreciar seu interesse e decidir sobre a vantagem ou não de incorporá-la ao conhecimento existente" (p. 16). Contudo, Brookes afirmou que a informação deve ser apresentada dentro de uma estrutura cognitiva relevante. Assim ele estava considerando, mesmo que subjetivamente, outros aspectos incluindo também os de natureza psicossomática.

Araujo (2014) alerta que as pesquisas desenvolvidas nas duas últimas décadas reformularam o conceito de conhecimento abordado pela equação de Brookes. Ou seja, não é mais a simples adição de dados a um estado mental, mas, é compatível com um "[...] quadro mais complexo relacionado com diferentes processos de assimilação, acomodação, interpretação, imaginação, análise e síntese [...]" e evidenciando "[...] o caráter essencialmente contextual e intersubjetivo dos fenômenos informacionais". Por outro lado, Araujo também enfatiza a importância do modelo cognitivo (na qual a equação de Brookes está inserida) como sendo complementar aos outros dois paradigmas apresentados na seção 2.1.2. Contudo, Brookes apenas não descreveu em detalhes o processo de modificação do estado mental. Mas, ele chamou atenção que a chegada da nova informação não causa automaticamente

crescimento do conhecimento, ou seja, não pode ser admitida como um simples incremento de informação, mas, depende do tipo de relação que ela tem com as estruturas cognitivas do sujeito.

Além disso, alertam Batista, Costa e Alvares (2007), a proposta de Brookes tem sido apenas parcialmente aceita porque a CI tem se dedicado à coleta e organização para uso dos registros do mundo 3 de Popper, e não estudo das interações dos mundos 2 e 3 como propõe Brookes em sua equação, sendo esses ocupados por uma área denominada de Gestão do Conhecimento. Dessa forma, os autores recomendam que seja feito um resgate da proposta de Brookes para ampliar a área de atuação da Gestão do Conhecimento. Contudo, Brookes enfatiza muito em seus artigos a ligação da CI com o relacionamento existente entre os mundos 2 e 3 de Popper, e também estabelece uma base para a sua equação que segue mais fortemente o mundo 2.

2.1.5 Considerações finais da seção

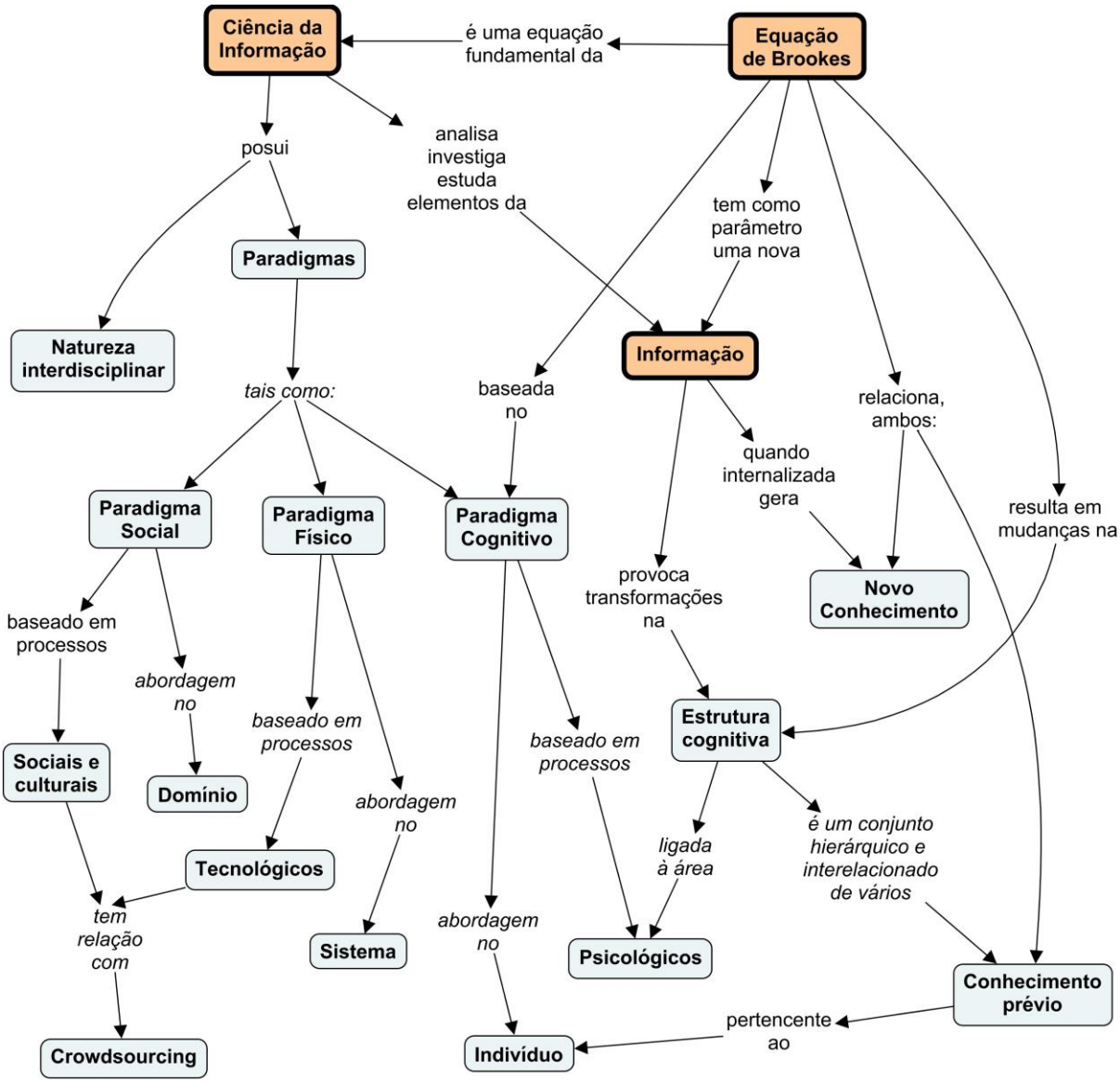
Apesar da CI ter uma dimensão social bem destacada ela está fortemente presente em diferentes áreas científicas, tecnológicas e culturais. Reforçando ainda mais essa abrangência da CI, ela tem uma natureza eminentemente interdisciplinar. Pinheiro (2007) destaca ainda que ela "[...] vai ser determinada e delimitada por essas relações interdisciplinares, em processo de constante mutação, como um organismo vivo" (p. 102). Porém, Moraes (2013), em sua análise sobre os caminhos da CI na contemporaneidade, alerta que o momento é de muita incerteza em vários dos aspectos da CI: histórico, paradigmas, formação e atuação profissional.

Analisando os paradigmas descritos por Capurro, percebe-se que existe uma complementação importante entre eles. O paradigma físico foca na tecnologia que é fundamental nos dias de hoje, devido às exigências cada vez maiores dos usuários (paradigma cognitivo), e também devido ao dilúvio informacional existente na sociedade (paradigma social) suportado por meios tecnológicos. O paradigma cognitivo é decisivo para estabelecer o aumento do conhecimento por parte do usuário, fortemente apoiado por meios tecnológicos (paradigma físico), na medida em que ele tem contato com novas informações e estas se relacionam com a sua bagagem cultural (paradigma social). O paradigma social revela-se também em movimentos cooperativos mais contemporâneos, tal como o *crowdsourcing*[7], para

[7] *Crowdsourcing* é uma prática que se utiliza da inteligência coletiva, geralmente de voluntários espalhados pela web, para resolver problemas, criar conteúdo e soluções, desenvolver novas tecnologias entre outros.

alimentar bases de dados considerando o usuário inserido nos seus contextos sociais com a informação dependente do domínio a qual ela pertence. Várias outras relações e dependências podem ser estabelecidas nesses três paradigmas que se completam para uma CI mais abrangente e interdisciplinar.

Figura 2 – Mapa conceitual com alguns relacionamentos abordados na seção 1: Ciencia da Informação



Fonte: Elaboração própria

Apesar das várias controvérsias sobre a equação fundamental da CI de Brookes, percebe-se a grande dimensão e contribuição dela para a CI tal como relatado por vários autores. A ideia básica da equação, de que uma nova informação, obtida pela motivação de

um desejo ou necessidade informacional do usuário, provoca transformações nas estruturas cognitivas desse usuário relacionando-se com o seu conhecimento prévio, é fundamental. É ainda importante destacar que esse processo básico é compatível com a Teoria da Aprendizagem Significativa que será abordada na seção 2.2.5.

O mapa conceitual da Figura 2 apresenta alguns relacionamentos importantes abordados nessa seção sobre Ciência da Informação, destacando, em cor alaranjada e espessura maior, alguns conceitos relevantes para a presente tese. Entre as várias proposições existentes no mapa, destacam-se aquelas que caracterizam os três paradigmas da CI: físico, social e cognitivo, relacionando este último com a equação de Brookes. Outras proposições revelam a formação de novo conhecimento por intermédio da informação internalizada, e o relacionam com o conhecimento prévio do indivíduo provocando alterações na sua estrutura cognitiva. Destaque também para o *crowdsourcing* que relaciona os paradigmas social e físico, funcionando por intermédio de um processo social e apoiado pela tecnologia.

2.2 Conceito, mapa conceitual e aprendizagem significativa

Essa seção aborda definições de conceito e a teoria do conceito de Dahlberg (1978), relacionamentos entre conceitos com representações por proposições e hipertexto com suas variações tais como hiperdocumento e hipermídia, caracteriza os mapas conceituais, apresenta a teoria da aprendizagem significativa, cita aplicações de mapas conceituais e trata da disseminação de informações por intermédio de mapas conceituais.

2.2.1 Conceito

A linguagem é usada como forma de expressão dos pensamentos humanos. Com ajuda dela o ser humano é capaz de relacionar-se com objetos (coisas, fenômenos, processos, acontecimentos etc.) ao seu redor como também elaborar enunciados sobre eles, chamados por Dahlberg (1978) de conceitos. A realidade simplificada, generalizada da realidade, adquirida mediante a existência e uso de conceitos, segundo Moreira e Masini (1982), “torna possível a invenção de uma linguagem com relativa uniformidade de significados para todos os membros de uma cultura facilitando a comunicação interpessoal” (p. 27). Dessa forma, os autores sinalizam que “o homem vive muito mais num mundo de conceitos do que de objetos, eventos e situações” (p. 27).

O processo de organizar o conhecimento em conceitos na maioria das vezes gira em torno de métodos dedutivos e indutivos. O método indutivo parte da representação dos elementos/objetos e relações imersos num contexto, e "[...] o método dedutivo propõe que se elaborem mecanismos de abstração para pensar primeiramente o domínio/contexto, independentemente de pensar os elementos e suas relações" (CAMPOS; SOUZA; CAMPOS, 2003, p. 10). Existem vários métodos para organização do conhecimento na literatura, analisados por Campos *et al.* (2003, p. 10), tais como a Teoria da Classificação Facetata de Ranganathan em 1951, que basicamente usa um modelo dedutivo; a Teoria do Conceito proposta por Ingetraut Dahlberg em 1978 propõe um modelo híbrido, ou seja, tanto dedutivo quanto indutivo; a Teoria da Terminologia de E. Wuester em 1981 suporta métodos indutivos; a Ontologia Formal de Nicola Guarino em 1998, que usa métodos indutivos; Edgar Morin em 2000 propôs um processo circular que vai da separação à ligação, da análise à síntese dos conceitos; Maria L. de A. Campos em 2001 sintetiza os modelos usados em dois tipos, dedutivo e indutivo.

Nesse trabalho será assumida como base teórica para o processo de organizar o conhecimento em conceitos a Teoria do Conceito, formulada por Ingetraut Dahlberg, que define conceito como um enunciado sobre objetos, fenômenos, processos, acontecimentos etc. (DAHLBERG, 1978). Numa visão geral, essa teoria afirma que os conceitos podem ser individuais ou gerais. Conceitos individuais são, por exemplo, "a UnB" e "o artigo científico de fulano". Enquanto conceitos gerais são, por exemplo, "as universidades" e "os artigos científicos", relativos aos exemplos de conceitos individuais anteriores. Além disso, pode-se formular enunciados sobre os conceitos onde cada enunciado verdadeiro representa elementos do respectivo conceito. Como exemplo, no conceito "a UnB" poder-se-ia formular os enunciados: "é uma instituição", "possui pós graduação em Ciência da Informação" etc. O conjunto dos anunciados verdadeiros sobre um determinado objeto, fornece o conceito sobre o mesmo.

Ainda segundo Dahlberg (1978), a análise de um conceito é feita pela sua decomposição através da formulação de enunciados verdadeiros para se obter suas características. Essas características podem ser hierarquizadas de tal forma que o predicado de um enunciado pode-se tornar o sujeito de um novo enunciado até a formação de uma categoria, que é o conceito na sua mais ampla extensão. Características podem ser classificadas em espécies, tais como, matéria, quantidade, relação, processo, modo de ser, passividade, posição, localização e tempo. Dahlberg tece ainda classificações para as características: essenciais (constitutivas da essência ou consecutivas da essência) ou

acidentais (gerais ou individualizantes). Além disso, existe uma categorização formal dos conceitos por intermédio dos elementos: objetivos, fenômenos, processos, propriedades, relações e dimensão. O autor ainda chama atenção que a junção desses elementos é dependente da língua.

2.2.2 Relacionamento entre conceitos

O relacionamento entre conceitos é um item recorrente em várias áreas do conhecimento. Nessa subseção ele é abordado no formato de proposições e hiperlinks. Em subseções mais adiante o relacionamento entre conceitos é também abordado enquanto frases de ligação que estabelecem as proposições de um mapa conceitual (subseção 2.2.3), dados ligados no contexto da Web Semântica (subseção 2.3.3) e conexões entre nós de rede (subseção 2.3.3).

2.2.2.1 Proposição

Ludwig Wittgenstein, filósofo do século 20, argumentou em seu trabalho que a linguagem é composta por proposições que também representam os pensamentos no mundo (BILETZKI; MATAR, 2014). Bertrand Russell, há cerca de 100 atrás, em seu trabalho "On propositions: what they are and how they mean" (RUSSELL, 1919), definiu proposição como uma frase que pode ser julgada, necessariamente, como sendo verdadeira ou falsa. Esse valor é dependente do contexto dos conceitos ao qual a proposição se refere. Por exemplo, a frase "a construção de mapas conceituais é fácil" pode ser verdadeira para um grupo de pessoas com experiência na construção de mapas conceituais, porém, falsa para outro grupo que não conhece mapas conceituais. Outros autores, contemporâneos, Silva, Finger e Melo (2006), Bispo, Castanheira, Souza Filho (2011) e Menezes (2013) seguem a mesma assertiva sobre a definição de proposição, ou seja, uma frase, uma sentença, um pensamento à qual se pode atribuir juízo. Na Lógica Matemática, esse juízo possui os valores possíveis: verdadeiro ou falso. Dessa forma, os termos 'a construção de mapas conceituais' e 'fácil' não são proposições, mas conceitos, pois não faz sentido atribuir-lhes um valor. Por outro lado, o seu relacionamento com o conector 'é', os tornaria uma proposição, 'a construção de mapas conceituais é fácil', podendo receber o valor verdadeiro ou falso.

De uma maneira geral, uma proposição conecta dois conceitos estabelecendo entre eles uma relação. Segundo a Teoria do Conceito de Dahlberg (DAHLBERG, 1978), se

conceitos diferentes possuem características idênticas, então existe relação entre eles. A relação entre dois conceitos pode ser de dois tipos:

- **Relação lógica**: identidade (as características dos conceitos são as mesmas), implicação (um conceito está contido no outro), intersecção (os dois conceitos tem características comuns), disjunção (nenhuma característica em comum entre os dois conceitos), negação (um conceito possui uma característica que é a negação da característica do outro conceito);
- **Relação semântica**:
  - **Hierárquica**: um conceito possui uma característica a mais do que o outro, nesse caso o primeiro é o mais específico e o segundo o mais genérico;
  - **Partitiva**: quando um conceito é formado por outros conceitos, então esses últimos são partes do primeiro;
  - **De oposição**: quando um conceito é declarado o oposto do outro;
  - **Funcional**: normalmente aplicada a conceitos que expressam processos.

O estabelecimento de relações entre conceitos é bastante natural e até necessário em vários processos e abstrações que representam informação e conhecimento. Por exemplo, Motta (1987) propõe um método relacional para a construção de tesauros, enquanto sistemas conceituais, que usa o estabelecimento de relações entre os seus termos tendo como pressuposto a Teoria do Conceito de Dahlberg. Almeida (1998) chama atenção para o fato de que, na área de Terminologia, os conceitos não estão isolados, mas, formam redes de relações entre si. A busca de relações entre termos é também um dos aspectos básicos na construção de tesauros (MOTTA, 1987). A construção de um mapa conceitual, abordada na subseção 2.2.3, é um processo de criação de significados que implica em elencar uma lista de conceitos e criar proposições formadas por esses conceitos relacionados entre si por intermédio de frases de ligação (NOVAK; GOWIN, 1984).

2.2.2.2 Hiperlink

Um hiperdocumento apresenta informações de forma não linear e onde o percurso da leitura ou acesso é escolhido e realizado pelo leitor, através das conexões ou *hiperlinks* existentes entre os vários nós informacionais. Hiperdocumentos podem agregar e oferecer ao usuário diferentes mídias e formas de comunicação para apresentar a informação, tais como

textos, imagens, sons, vídeos etc. Por isso, ele também é chamado de hipermídia. Historicamente o termo originou-se da palavra hipertexto pelo fato de, na época, existirem predominantemente textos para apresentação da informação no meio digital. Porém, ainda hoje, é comum usar o termo hipertexto para designar hiperdocumentos ou hipermídias, sendo ainda que, em todos os casos, o relacionamento entre os seus nós com informações denomina-se *hiperlink*.

Hiperobjeto, termo mais recente, agrega conjuntos de informações advindos de diversas mídias, tal como hipermídia, porém, adiciona objetos compostos de ações ou funções como nós que também são interligados entre si por meio de hiperlinks. Como exemplo de hiperobjeto, um eletrodoméstico, "[...] quando informações tais como um manual de usuário, rede de assistência técnica, lojas de peças e acessórios podem ser facilmente acessados, seja por hiperlinks presentes no objeto físico - como códigos de barra, QR Codes - ou através de representações digitais interativas como realidade aumentada" (PEZZI, 2015, p. 176). Pezzi enfatiza para o fato de que "[...] aplicações científicas e educacionais de hiperobjetos são modelos ideais para estes, em que a omissão e o obscurecimento de informações não são desejados" (p. 177). Nessa linha, os hiperinstrumentos científicos ou educacionais são instrumentos cujas "[...] representações digitais contêm detalhes que facilitem, a qualquer pessoa interessada, aprofundar seus conhecimentos nos diversos aspectos do instrumento, de modo a garantir o seu uso, estudo, reprodução, adaptação e disseminação" (PEZZI, 2015, p. 195).

O desenvolvimento de hipertextos já é antigo, como mostram algumas situações ao longo da história. Por exemplo, a Enciclopédia de Diderot e D'Allembert, iniciada no século XVIII, com 35 volumes foi o foi o primeiro livro a ser criado no formato de hipertexto (PARENTE, 1999). Isso porque determinadas expressões da enciclopédia remetiam a outras partes da obra. Vannevar Bush, em 1945, já anunciava a ideia de hipertexto quando criticou que a maior parte dos sistemas de indexação da época era hierárquica ao invés de ser por associações tal como a mente humana funciona (BUSH, 1945). Parente (1999) também destaca que, em 1965, Theodore Nelson inventou o conceito de hipertexto para "[...] exprimir a ideia de um texto de dimensões cósmicas, informatizado, contendo todos os livros, incluindo imagens e sons, acessível à distância e navegável de forma não linear" (p. 73). Porém, antes de tudo, a leitura de uma enciclopédia clássica já é considerada hipertextual, pois usa ferramentas de orientação tais como léxicos, índices, tesauros, atlas, quadros de sinais, sumários e remissões ao final dos artigos (LÉVY, 1996).

Um hipertexto pode ser encarado como um espaço para leituras possíveis, ou seja, um texto é uma leitura particular de um hipertexto que o leitor leu seguindo um determinado caminho, a partir da matriz de textos potenciais que o autor do hipertexto construiu (LEVY; COSTA, 1999). Nessa mesma linha e no contexto da CI, hipertexto é considerado um "[...] complexo sistema de estruturação e recuperação de informação de forma multissensorial, dinâmica e interativa" (PARENTE, 1999, p. 80). Segundo Lévy (1993), o hipertexto possui seis características básicas: (i) metamorfose, por ele encontrar-se em constante construção e renegociação; (ii) heterogeneidade, pois os seus nós podem ser compostos por palavras, sons, imagens etc.; (iii) multiplicidade e encaixe das escalas, porque os seus nós ou conexões, podem ser eles mesmos uma rede de nós e conexões, sucessivamente; (iv) exterioridade, pelo fato de que o crescimento e diminuição da rede, bem como sua composição e recomposição, dependem da adição ou subtração exterior de elementos ou conexões; (v) topologia, devido a seu funcionamento ocorrer por proximidade; (vi) mobilidade dos centros, pois os vários centros da rede são móveis, formando ao redor de si uma ramificação em estrutura de rizoma, sem raiz ou hierarquia, tipo uma vegetação que flutua na água.

Em termos de ferramentas de criação de hipertexto, elas já existem há bastante tempo e de forma bem acessível como, por exemplo, um simples e popular software editor de textos, tal como relata Cristovão (2000) em uma experiência realizada em contexto educacional. Hoje a tecnologia informática continua sendo um grande aliado na criação de hipertextos. Contudo, Campos, Souza e Campos (2003) sinalizam em seu trabalho alguns problemas na sua construção, principalmente no âmbito da comunicação entre o autor que desenvolve o conteúdo temático, e o analista de sistemas por intermédio da falta de metodologias apropriadas para a implementação de sistemas que possam representar unidades do conhecimento.

Ainda sobre o processo de autoria de hipertexto, Lima (2005) destaca que ele é, em última análise, um processo de classificação onde o "autor planeja a estrutura global do hipertexto, seleciona símbolos apropriados (i.e., palavras, ícones) e cria links eletrônicos para representá-los" (p. 2). Isso ocorre de forma similar ao processo de classificação de documentos onde o "classificador determina o conteúdo, seleciona termos apropriados e cria pontos de acesso" (p. 2).

2.2.3 Caracterização dos mapas conceituais

Os mapas conceituais foram desenvolvidos nos anos 70 por Joseph Donald Novak, professor da Cornell University e pesquisador no Florida Institute for Human & Machine Cognition, e apresentados na sua publicação “A Theory of Education” (NOVAK, 1977). Eles foram baseados na Teoria da Aprendizagem Significativa de David Ausubel (AUSUBEL, 1968) e são estudados na da área da Ciência Cognitiva (SHERRATT; SCHLABACH, 1990).

Mapas conceituais são ferramentas para organização, representação e compartilhamento de conhecimento (CAÑAS et al., 2004). Sherrat e Schlabach (1990) definem mapa conceitual como elementos capazes de envolver a identificação de conceitos ou ideias pertencentes a um assunto e a descrição das relações que existem entre essas ideias na forma de uma descrição esquemática, onde o sua finalidade é representar a compreensão de um indivíduo sobre um corpo de conhecimento e ilustrar o relacionamento entre as ideias significativas para ele. Podem também serem vistos como “diagramas hierárquicos que procuram refletir a organização conceitual de uma disciplina ou parte de uma disciplina” (MOREIRA; MASINI, 1982, p. 45). Ou, simplesmente, como uma técnica para representação gráfica de conhecimento (LANZING, 1997).

Sobre a sua representação gráfica, os mapas conceituais consistem de um conjunto de conceitos, cada um dentro de uma caixa, e relacionamentos entre conceitos indicados por linhas de conexão com a sua respectiva descrição. No contexto do mapeamento de padrões semânticos que são definidos por códigos gráficos, colocar conceitos dentro de caixas contornadas vai ao encontro do que Ware (2010) argumenta como uma boa solução. Da mesma forma esse autor também indica um bom simbolismo empregado para relacionar duas entidades, ou dois conceitos, como sendo uma linha que é também usada nas ligações entre dois conceitos de um mapa conceitual.

Novak e Gowin (1984) definem o termo conceito como uma regularidade que acontece em eventos ou em objetos que são designados por um termo. Dessa forma, ‘vento’ é um termo usado para o evento que representa o ar em movimento enquanto o termo ‘cadeira’ representa um objeto com características tais como possuir pernas, assento e costas. Porém, os autores ainda afirmam que é necessário conhecer uma linguagem para usar essas designações. Essa definição de conceito de Novak e Gowin, no contexto dos mapas conceituais, é compatível com a definição de Dahlberg (1978) na seção 2.2.1, quando afirma que conceitos são enunciados sobre objetos, fenômenos, processos, acontecimentos etc.

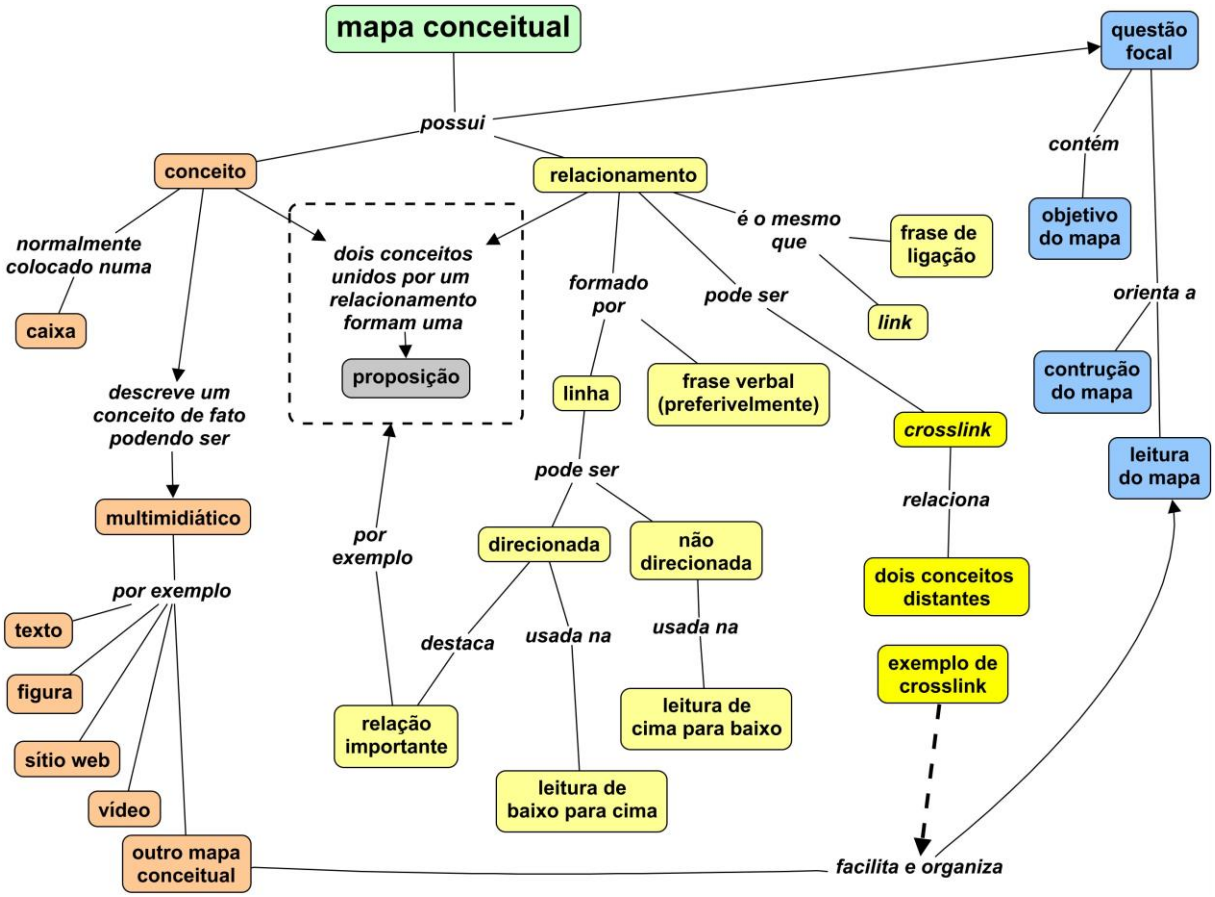
São elementos fundamentais de um mapa conceitual:

- **Conceito**: representa um conceito de fato, porém, não necessariamente textual. Ele pode ser multimidiático, isto é, figura, vídeo, sítio da web etc. São, predominantemente, escritos no interior de uma caixa.
- **Relacionamento, frase de ligação ou *link***: faz a ligação entre dois conceitos. Normalmente é composta por verbo, pois facilita a leitura e o entendimento de forma independente a cada dois conceitos ligados. São representados por uma linha com a descrição do relacionamento dos conceitos aos quais ela conecta. A indicação da direção do relacionamento, por intermédio de uma seta na linha, facilita a sua leitura. Contudo, é também comum não colocar seta quando a ordem de leitura ocorre naturalmente de cima para baixo. Ainda sim, em alguns casos, usa-se uma seta para destacar a ordem de leitura de um relacionamento.
- **Proposição**: é o conjunto de dois conceitos e uma frase de ligação. É desejável que a proposição faça sentido completo, ou que se comporte como uma proposição de Bertrand Russell, tal como abordada na subseção 2.2.2.1, ou seja, capaz de assumir um valor verdadeiro ou falso;
- **Questão focal**: é uma questão, ou uma pergunta, que representa o tema ou o escopo do mapa conceitual. Tem o objetivo de direcionar e orientar a construção do mapa para que o aprendiz não se desvie do foco inicialmente pretendido;
- ***Crosslink***: é uma frase de ligação que conecta conceitos distantes no mapa. Criar *closslinks* é desafiador para o autor do mapa, pois, normalmente, eles estabelecem a integração entre conceitos ou subdomínios pertencentes a áreas distintas do conhecimento.

A Figura 3 apresenta um mapa conceitual com a questão focal ‘Quais são os elementos fundamentais de um mapa conceitual?’. A preocupação com a legibilidade do mapa é importante e deve ser trabalhada por intermédio de elementos de formatação. Nesse mapa, da Figura 3, foram usadas cores para identificar grupos de conceitos pertencentes aos subconjuntos de: ‘conceito’, ‘relacionamento’, ‘proposição’, ‘*crosslink’* e ‘questão focal’. A estrutura do mapa é flexível e capaz de absorver a criatividade do autor para representar uma organização da informação e formatações que facilitem a leitura e entendimento do assunto abordado. Por exemplo, a parte que explica o que é uma proposição, e o exemplo de *crosslink,*

não seguem ao padrão de proposições normalmente encontradas em mapas conceituais, isto é, dois conceitos interligados por uma frase de ligação.

Figura 3 – Mapa conceitual com a questão focal: ‘Quais são os elementos fundamentais de um mapa conceitual?’



Fonte: Elaboração própria

Sobre representações que se parecem com mapas conceituais, Gaines e Shaw (1995) constataram, em sua investigação, que o termo mapa conceitual é usado na literatura para referenciar uma grande variedade de representações gráficas de conhecimento, tais como:

- **Mapas conceituais hipermídia**: na medida em que os mapas conceituais possibilitam o trabalho com cores, mídias, navegação etc. podendo oferecer uma boa interface para sistemas de hipertexto ou hipermídia;
- **Mapas conceituais formais**: no geral, mapas conceituais tendem a ser mais formais e exigir mais restrições em sua construção do que outros diagramas;
- **Redes semânticas**: uma vez que a representação de bases de conhecimentos de redes semânticas, usando uma linguagem formal, não é de fácil compreensão, os

mapas conceituais podem servir como um bom meio para a visualização e edição de redes semânticas.

Por outro lado existem representações conceituais que se diferem muito dos mapas conceituais, tal como o mapa mental. Nesse caso, a principal diferença é que enquanto o mapa conceitual possui frases de ligação que definem proposições, o mapa mental se utiliza de uma ligação não nomeada entre conceitos ou descrições que, normalmente, indica uma hierarquização ou uma relação de composição entre eles. Moreira (2013) alerta que às vezes os mapas conceituais são confundidos com esquemas, diagramas organizacionais ou organogramas. Moreira ainda sinaliza que mapas conceituais não são guiados por temporalidade, direcionalidade ou hierarquias organizacionais ou de poder, como ocorre em outros tipos de diagramas.

2.2.4 Construção e avaliação de mapas conceituais

Especificamente sobre o passo-a-passo de como construir um mapa conceitual existem sugestões tais como as encontradas nos estudos de Novak (2010) com 10 passos e Moreira (2013) com 11 passos, embora ambos autores enfatizem que dado um conjunto inicial de conceitos os mapas resultantes será diferente para cada aprendiz que construí-lo, mesmo seguindo o roteiro sugerido.

Outro aspecto relevante é que "contrariamente a textos *et al.* materiais instrucionais, os mapas conceituais não dispensam explicações do professor" (MOREIRA; MASINI, 1982, p. 50). Por conta dessa observação, a preocupação quanto à legibilidade de um mapa conceitual se torna um dos focos principais em sua construção. Uma das indicações mais importantes é quanto ao estabelecimento de relações significativas. Moreira e Masini (1982, p. 50) afirmam que "o agrupamento de conceitos em combinações potencialmente significativas é responsável pela formulação e entendimento de proposições". Ter em mente a forma como eles podem ser avaliados já é um bom indicativo para a sua construção. Cañas, Novak e Reiska (2015) destacam que um mapa conceitual pode ser avaliado segundo critérios estruturais e de conteúdo, porém o propósito do mapa conceitual é determinante para que a avaliação seja justa. De um modo geral, eles elencaram uma lista de critérios para obtenção de um bom mapa:

- Estabelecer uma questão focal para que o contexto do mapa seja bem definido e explicitado;

- Conceitos devem ser nomeados com uma ou poucas palavras;
- Frases de ligação devem ser formadas por uma ou poucas palavras e não devem conter conceitos relevantes para o mapa;
- Mapas conceituais devem possuir uma organização hierárquica com os conceitos mais gerais em cima e os mais específicos em baixo;
- No geral, não mais do que 3 ou 4 conceitos devem ser ligados a um outro conceito;
- *Crosslinks* devem especificar relacionamentos significativos entre dois conceitos em diferentes subdomínios mostrados no mapa. Eles são mais facilmente colocados quando o mapa está próximo do fim;
- Um conceito não deve aparecer mais do que uma vez no mapa conceitual.

Cañas, Novak e Reiska, ainda alertam que apesar desses critérios serem necessários, eles não são suficientes para garantir que um mapa seja considerado bom, pois é o mesmo que ocorre com um bom poema que tem as razões para sua boa qualificação como difíceis de serem explicitadas. Porém, uma pessoa com experiência na construção de mapas conceituais saberia identificar um bom mapa mesmo que ele não saiba explicar as razões para essa decisão.

É preciso ter cuidado na hora de avaliar o mapa de alguém, pois, o mesmo mapa pode tanto significar uma profunda má interpretação por parte do aluno quanto significar um modo criativo não usual de representar as relações conceituais (NOVAK; GOWIN, 1984). Além disso, um mapa poder ser interpretado de várias formas, pois “[...] análises recentes mostram que toda leitura modifica seu objeto e que (como Borges já dizia) uma literatura se difere de outra menos pelo seu texto do que pela forma como é lida" (PARENTE, 1999, p. 90). Portanto a avaliação de um mapa é um processo subjetivo e que devem ser levados em consideração vários fatores, tanto do lado do leitor/avaliador quanto do lado do autor. Além disso, um mapa conceitual nunca é prefeito.

Mapas conceituais são, sobretudo, estruturas para representar conhecimento, dessa forma, há flexibilidade quanto a formatações, hierarquia entre outras indicações. Por exemplo, Dutra, Fagundes e Cañas (2004) argumentam que são as frases de ligação que estabelecem as fronteiras de um conceito no mapa, segundo a visão Piagetiana de seu trabalho. Dessa forma, na visão desses autores, a hierarquia deixa de ser importante em um mapa conceitual, uma vez

que o autor do mapa pode ainda não ter o sistema hierárquico formado em sua mente e assim não ter elementos mínimos para essa demanda

2.2.5 Aprendizagem significativa

A Teoria da Aprendizagem Significativa foi concebida nos anos 60 por David Ausubel (AUSUBEL, 1968) e é baseada no modelo construtivista do processo cognitivo humano. De uma maneira geral, essa teoria defende a ideia de que novos conhecimentos são melhores aprendidos por um indivíduo quando eles são significativos, ou seja, conseguem se relacionar com o conhecimento anterior do aprendiz.

Ausubel (2000) argumenta que o 'conhecimento' já é significativo por definição. Ele é o produto significativo de um processo psicológico e cognitivo que envolve a interação entre ideias significativas e ideias já conhecidas e relevantes na estrutura cognitiva particular dos alunos e seu conjunto mental para aprender significativamente ou adquirir e reter conhecimento. Nessa teoria, existem conceitos fundamentais (AUSUBEL, 1968, 2000; MOREIRA; MASINI, 1982; MOREIRA, 2013):

- **Subsunçor**: é o conhecimento prévio que o aprendiz possui e que está estabelecido em sua estrutura cognitiva, servindo para se relacionar, ancorar e dar significado aos novos conhecimentos.
- **Estrutura cognitiva**: é um conjunto hierárquico de subsunçores relacionados entre si. Essa estrutura muda dinamicamente na medida em que novos conhecimentos se relacionam com os seus subsunçores.
- **Diferenciação progressiva**: é o processo de atribuição de novos significados a um subsunçor. Na medida em que um subsunçor se relaciona com novos conhecimentos ele se modifica e passa a ter novo significado.
- **Reconciliação integradora**: é a modificação da estrutura cognitiva, por intermédio dos outros subsunçores que estão relacionados ao subsunçor que sofreu uma diferenciação progressiva. Esse processo resolve inconsistências, elimina diferenças aparentes, integra significados, modifica hierarquias etc.

Segundo Ausubel, a aprendizagem significativa requer três condições básicas para acontecer:

- Um material instrucional claro com exemplos e linguagem relacionadas e próximas do conhecimento prévio do aluno;

- O aprendiz deve possuir subsunçores que estejam relacionados com o novo conteúdo que será abordado;
- É necessário que o aluno esteja com vontade de aprender nesse formato da aprendizagem significativa usufruindo-se da diferenciação progressiva e da reconciliação integradora.

Novak (2010, 2011), com sua teoria da educação baseada na aprendizagem significativa de Ausubel e na epistemologia construtivista (que estabelece relações fortes entre os elementos professor, aprendiz, matéria de ensino, contexto e avaliação), resume em seis os princípios fundamentais que servem de guia básico para facilitar quem estiver envolvido em processos de ensino-aprendizagem:

(i) Deve haver motivação para aprender. Nenhum aprendizado ocorrerá a menos que o aprendiz decida aprender;

(ii) Deve-se entender e aproveitar o conhecimento prévio e relevante do aprendiz, considerando ideias válidas e inválidas;

(iii) Deve-se organizar o conhecimento conceitual que se quer ensinar;

(iv) O aprendizado ocorre em um contexto, e deve-se levá-lo em consideração como um facilitador para a educação;

(v) A aprendizagem pode ser auxiliada por um professor que é experiente e sensível às ideias e sentimentos do aluno; e

(vi) Avaliações são necessárias para medir o progresso e motivar o aluno.

Novak ainda acrescenta que esses princípios servem tanto para instituições educacionais quanto para empresas de uma maneira geral. Isso porque ele alerta que mesmo as empresas estão se movendo na direção de enxergar os seus clientes como potenciais professores e aprendizes. Está se chegando a um consenso de que os processos de aprendizagem significativa são os mesmos operados por cientistas, matemáticos ou especialistas de qualquer área, quando esses elaboram conhecimento novo (NOVAK; CAÑAS, 2007, 2008). A abrangência dessa teoria é grande e tem feito contribuições significativas em várias áreas do conhecimento.

Novak e Gowin (1984) usam a Teoria da Aprendizagem Significativa para avaliar mapas conceituais, de acordo com os três princípios:

(i) Se a estrutura cognitiva do mapa é organizada hierarquicamente com conceitos mais específicos subordinados aos conceitos e as proposições mais gerais e abrangentes;

(ii) Se os conceitos estão numa diferenciação progressiva, ou seja, acompanhados do reconhecimento de uma maior abrangência e especificidades nas regularidades dos objetos e acontecimentos; e

(iii) Se ocorre a reconciliação integradora, onde dois ou mais conceitos estão relacionados em termos de novos significados entre conceitos.

Novak e Gowin ainda enfatizam que um dos princípios mais importantes da educação, que é também corroborado por Ausubel (1968), é conhecer o que o aprendiz já sabe. Apesar disso não ser fácil de ser comprovado, é possível que com os mapas conceituais haja mais possibilidades de se chegar a esse conhecimento prévio do aprendiz.

2.2.6 Aplicações de mapas conceituais

No contexto da Ciência da Informação (CI), Sherrat e Schlabach (1990) defendem a ideia de que, tanto na formação de profissionais da CI, no lado do aluno e do professor, quanto na prática e na capacitação do profissional em serviço, os mapas conceituais podem ser úteis. Mais recentemente, e especificamente na Biblioteconomia, Rodrigues e Cervantes (2015) sugeriram o uso de mapas conceituais como processo na organização e representação do conhecimento, com o objetivo de melhorar o tratamento temático da informação e propiciar um subsídio intelectual por meio de analogias. De fato, como sinalizavam Novak e Gowin (NOVAK; GOWIN, 1984), estratégias espaciais são necessárias para ajudar as pessoas na aquisição, armazenamento, reestruturação, comunicação e utilização de conhecimento, bem como para superar as limitações da capacidade de memória.

Tergan (2003) apresenta e justifica o uso de mapas conceituais para o gerenciamento da informação e do conhecimento. Belluzzo (2006) também discute o uso de mapas conceituais na gerência da informação e produção do conhecimento e coloca a seguinte questão: “que abordagem podemos utilizar para desenvolver um conjunto de atitudes e condutas que possam auxiliar no uso e domínio da informação?” (p. 86). A autora sugere como resposta o uso da aprendizagem significativa por entender que ela auxilia a pensar e manter conexões entre conceitos e sua estrutura além de facilitar as relações ente diferentes campos do conhecimento.

Lanzing (1997) relaciona alguns propósitos para os mapas conceituais:

(i) Na criação de ideias (fazendo um *brain stormin*);

(ii) Para projetar uma estrutura complexa (um texto muito longo, uma hipermídia, um grande website);

(iii) Para apresentar ideias complexas;

(iv) Para ajudar no aprendizado pela integração explicita de novos e antigos conhecimentos; e

(v) Para avaliar o aprendizado ou diagnosticar conhecimentos mal-entendidos.

A educação é uma das áreas mais fortes de aplicação de mapas conceituais, onde "o professor e o aluno podem usufruir da potencialidade dos mapas conceituais para o alcance de diversos objetivos: apresentar, avaliar, aprender, organizar, integrar, sintetizar, refletir etc." (CRISTOVÃO *et al.*, 2014). A produção de mapas conceituais auxilia alunos a aprenderem, pesquisadores a elaborarem novos conhecimentos, administradores a estruturarem e gerenciarem empresas, escritores a escreverem melhor e professores a avaliarem o aprendizado (NOVAK; CAÑAS, 2008).

Novak e Cañas (2007, 2008) propõem um novo modelo de educação baseado em mapas conceituais e na aprendizagem significativa de Ausubel (1968). Com o uso do software CmapTools[8], desenvolvido pelo Florida Institute for Human & Machine Cognition (IHMC)[9], torna-se disponível uma grande variedade de funções para operação com mapas conceituais, desde a criação e edição de mapas com gravação do histórico de ações, até a formatação completa, o trabalho colaborativo e o compartilhamento dos mesmos por meio de servidores públicos ou do próprio IHMC. O CmapTools também é considerado um ambiente de software para compartilhamento e modelagem de conhecimento (CAÑAS et al., 2005).

Há diversas experiências de mapas conceituais na área de educação relatadas em artigos e eventos. Por exemplo, Williams (1998) desenvolveu um trabalho com os mapas conceituais para avaliar o quanto alunos da educação superior na área da Matemática compreenderam a matéria. Nesse estudo o autor conseguiu classificar níveis do entendimento conceitual dos alunos, a partir da comparação de seus mapas e com os mapas de especialistas na área. Novak e Cañas (2008) sugerem duas técnicas de aplicação de mapas conceituais na educação:

[8] Software CmapTools, disponpivel em: <http://cmap.ihmc.us/>.
[9] IHMC, acesso em: <http://www.ihmc.us/>.

(i) Estacionamento: é sugerido ao aluno um grupo de conceitos como ponto de partida para a construção do mapa; e

(ii) Esqueleto: é sugerido um mapa com conceitos e proposições, previamente e cuidadosamente, escolhidos de forma a dar condições à continuidade da construção por parte do aluno.

Novak e Cañas ainda sugerem o uso simultâneo das duas técnicas, pois os alunos poderão ampliar e aprimorar o mapa à medida que eles realizam atividades relacionadas ao assunto e aumentam sua compreensão gerando modelos de conhecimento complexos que interligam fontes, resultados e experimentos.

Finalmente, mapas conceituais são instrumentos que podem levar a "[...] profundas modificações na maneira de ensinar, de avaliar e de aprender" (MOREIRA, 2013, p. 48). Segundo o Moreira, eles promovem a aprendizagem significativa e o seu uso completo implica "[...] atribuir novos significados aos conceitos de ensino, aprendizagem e avaliação. Por isso mesmo, apesar de se encontrar trabalhos na literatura ainda nos anos setenta, até hoje o uso de mapas conceituais não se incorporou à rotina das salas de aula" (p. 48). Além disso, devido ao seu apelo visual, o mapa conceitual vai ao encontro do "funcionamento em estéreo do cérebro", termo cunhado no início dos anos 80 por Babin e Kouloumdjian (1983), que denota uma mixagem das aptidões cerebrais do lado esquerdo com o lado direito do cérebro. No caso dos mapas conceituais, pode-se integrar a competência da linguagem, pertencente ao lado esquerdo do cérebro, com a competência da espacialidade, nesse caso a disposição dos conceitos no mapa, representando o lado direito do cérebro.

2.2.7 Disseminação de informações pelos mapas conceituais

Um mapa conceitual que tem como objetivo apresentar ou disseminar um determinado assunto pode ser mais fácil de ser entendido do que essa mesma informação em formato textual. Porém, Cañas *et al.* (2015) destacam que mapas conceituais com propósito de comunicar algo devem ser claros e de fácil leitura. Um mapa conceitual pode ser considerado um diagrama, e apesar de sua simbologia gráfica simples, ele pode representar ideias e conhecimentos complexos. Diagramas são usados para planejar, projetar coisas e estruturar ideias, em resumo, "diagramas são ideias concretizadas" (WARE, 2010).

Diversos outros autores argumentam o quanto os mapas conceituais são bons para comunicar e disseminar informações. Vekiri (2002) argumenta que há menor esforço mental

para compreender um texto quando este está acompanhado de mapas conceituais. Orrantia (2012) mostra que os mapas auxiliam na disseminação da informação. Valerio, Leake e Cañas (2012) comprovam que mapas melhoram substancialmente as habilidades de compreensão de leitura dos usuários no quesito velocidade, em comparação à leitura exclusiva de texto. Lima (2004) argumenta que a característica gráfica do mapa conceitual auxilia a compreensão das relações entre os conceitos e do conhecimento no todo. De fato, Zhang (2008) observa que sem o auxílio de visualização gráfica, há necessidade de maior abstração de informações e, consequentemente, menor percepção ou compreensão dos dados e informações.

Belluzzo (2006) discute o uso de mapas conceituais enquanto instrumentos de apoio à gestão da informação e da comunicação e da mediação do desenvolvimento das habilidades de acesso e uso da informação na sociedade contemporânea. Belluzzo destaca, em especial, as áreas de informação e comunicação, e enfatiza o uso de práticas sociais para "[...] orientar a produção e o compartilhamento do conhecimento individual e coletivo a fim de atender às demandas por mediação dos novos instrumentos informáticos ou tecnologias de representação e comunicação dominantes no contexto atual e sua ágil inserção no cotidiano das pessoas".

Pesquisadores usam mapas conceituais para apresentar informações mais gerais de suas pesquisas, tendo como benefícios a exploração de questões em tempo real, criando um diálogo visual, e documentando os problemas com rapidez (CONCEIÇÃO; SAMUEL; BINIECKI, 2014). Esses autores ainda destacam que os mapas conceituais oferecem possibilidades para sintetizar ideias para orientar a análise e o processo visual permitindo que novas ideias emerjam. Destacam também o compartilhamento de mapas em servidores computacionais como processo de disseminação, seja entre pesquisadores, aprendizes ou usuários no geral.

2.2.8 Considerações finais da seção

Considerando a definição de conceito de Dahlberg, na seção 2.2.1, cada nó informacional de um hipertexto pode ser considerado também um conceito. Além disso, a ligação entre dois nós de um hipertexto pode ser equiparada a uma proposição simples, sem uma frase de ligação explícita, pois as ligações estariam quase sempre estabelecendo uma relação lógica de identidade. Todavia, a liberdade de autoria de quem cria um hipertexto pode levar a outros tipos de relações lógicas ou até mesmo relações semânticas. Ainda sim, há de se considerar que relações diferentes da identidade podem trazer confusão para o leitor. De

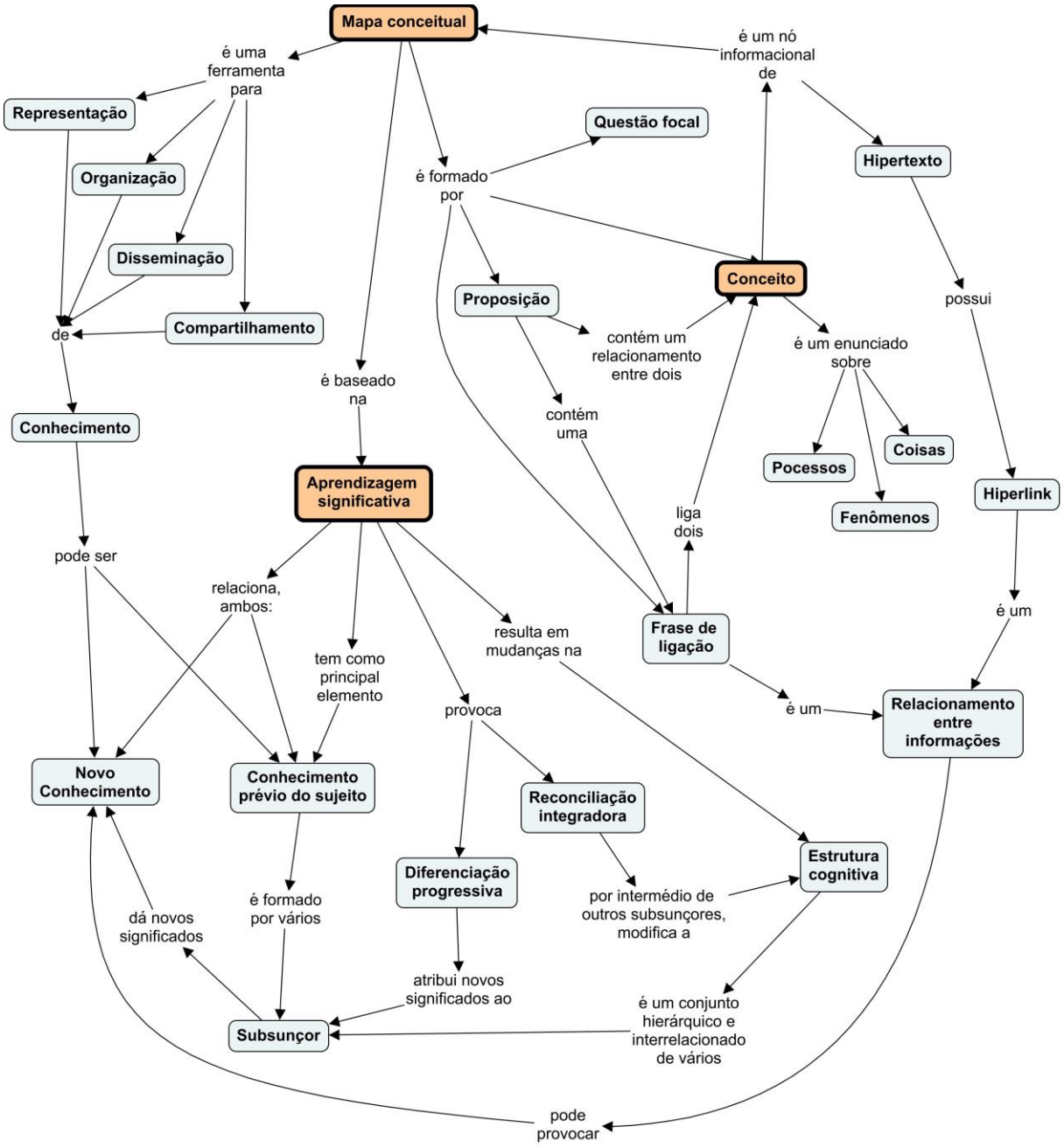
qualquer forma, um hipertexto é um conjunto de proposições ou uma rede de informação, tal como será abordado na seção 2.5, ou ainda pode ser considerado um mapa conceitual, como foi visto na seção 2.2.

Mapas conceituais já são usados desde os anos 70, porém, somente recentemente eles foram potencializados com o advento das ferramentas digitais para a sua construção, edição e disseminação. Eles têm grande abrangência em várias áreas do conhecimento, pois trabalham, sobretudo, com a representação de conhecimento. Graças à aprendizagem significativa eles são usados mais fortemente na educação por vários motivos, desde a facilidade de verificar o conhecimento prévio do aluno, passando pelo auxílio na construção do conhecimento pela modificação de suas estruturas mentais, até serem usados como ferramentas para auxílio à avaliação de aprendizagem. Apesar da ação de construir mapas conceituais ter um alcance mais profundo quando se trata da aprendizagem significativa, a disseminação de informações e processos de comunicação, por intermédio de mapas conceituais, também têm sido amplamente usados e foram verificados como ações possíveis por vários pesquisadores.

Atualmente eles são confundidos com os mapas mentais que são mais populares. Contudo, estes últimos são estruturas estritamente hierarquizadas, em forma de árvore, e não em forma de rede ou grafo. Além disso, eles não têm as importantes frases de ligação que estabelecem ricas proposições no caso dos mapas conceituais, tanto para o processo de construção quanto para o processo de disseminação ou comunicação da informação.

O mapa conceitual da Figura 4 apresenta alguns relacionamentos importantes abordados nessa seção sobre conceito, mapa conceitual e aprendizagem significativa, destacando, em cor alaranjada e espessura maior, alguns conceitos relevantes para a presente tese. Entre as várias proposições existentes no mapa, destacam-se os usos e a formação de um mapa conceitual, os relacionamentos entre a aprendizagm significativa com o novo conhecimento e o conhecimento prévio do sujeito, provocando a diferenciação progressiva, a reconciliação integradora bem como as alterações na estrutura cognitiva. Também se destaca o relacionamento entre informações caracterizado pelas frases de ligação de um mapa conceitual e dos hiperlinks de um hipertexto. Além disso, no contexto do presente trabalho, o mapa conceitual é uma boa ferramenta para representar, organizar, disseminar e compartilhar conhecimento.

Figura 4 – Mapa conceitual com alguns relacionamentos abordados na seção 2: conceito, mapa conceitual e aprendizagem significativa



Fonte: Elaboração própria

2.3 Web Semântica

A Web Semântica não é uma web[10] separada, mas uma extensão do atual modelo de web, onde a informação é associada a um significado e disponibilizada para acesso e trabalho conjunto entre computadores e pessoas (BERNERS-LEE *et al.*, 2001). Tem-se tentado associá-la ao termo web 3.0, como um próximo movimento da internet depois da web 2.0[11]. Além disso, Web Semântica e Web de Dados tem o mesmo significado (BIZER; HEATH; BERNERS-LEE, 2009; HEATH, 2009; KÉPÉKLIAN; CURÉ; BIHANIC, 2015). Percebe-se que, na literatura, alguns autores preferem o primeiro termo e outros o segundo.

Essa seção aborda a trajetória da Web Semântica, iniciando com a explanação das necessidades que motivaram a sua criação a partir da web. Também são apresentados os conceitos e aplicações de dados abertos, dados ligados e os dados abertos ligados. São abordadas as ontologias no contexto dos dados ligados e um exemplo de base de dados abertos ligados: a DBpedia.

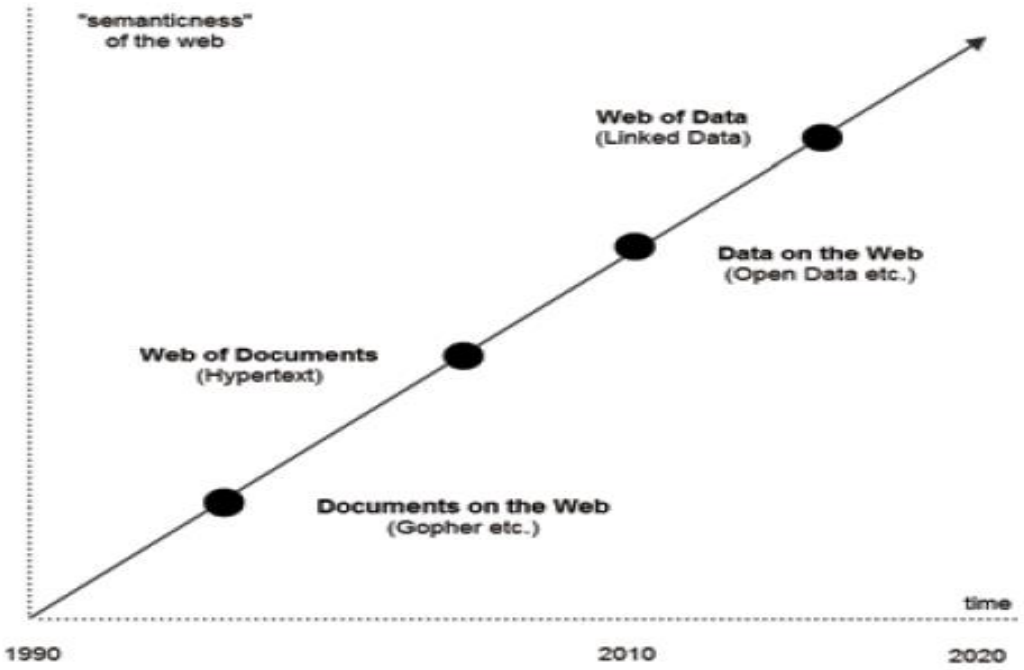
2.3.1 Da web a Web de Dados

A biblioteca integral sempre foi um sonho da humanidade e o projeto que mais se aproxima disso é a web (PARENTE, 1999). A propriedade essencial da web é a sua universalidade favorecendo a comunicação social, unindo línguas e culturas diferentes (BERNERS-LEE *et al.*, 2001). Ela tem características fortes tais como: é distribuída, tem grande volume de informações disponíveis, repositório não estruturado, é ubíqua e sofre constantes mudanças (BAEZA-YATES; RIBEIRO-NETO, 2011). Contudo, a web é muito mais do que um conjunto de partes tecnológicas (SHADBOLT et al., 2013), e a sua evolução tem sido concentrada na sua capacidade semântica. Bauer e Katenböck (2012) mostram, no Gráfico 1, a evolução da web desde o nascimento nos anos 90 com documentos sobre a web (*Documents on the Web*) - forma mais rígida de navegação onde o usuário indicava apenas um endereço para acesso direto e único a um documento; passando pela web de documentos (*Web of Documents*) – uma típica navegação pela web; depois, dados sobre a web (*Data on the Web*) – acesso a dados abertos na web e, finalmente, tendendo para o ano 2020, a Web de Dados (*Web of Data*) – conhecida também por Web Semântica (como já dito anteriormente) onde, tipicamente, os dados estão ligados e também podem ser interpretados por máquinas.

[10] Web é abreviatura universalmente usada para World Wide Web.
[11] Web 2.0 enfatiza a interação e colaboração entre usuários com serviços tais como, redes sociais, blogs, wikis, folksonomias, sites de compartilhamento de vídeos etc.

Resumidamente, o gráfico mostra a evolução da capacidade semântica da web (*semanticness of the web*).

Gráfico 1 – Evolução da capacidade semântica da web



Fonte: Bauer e Katenböck (2012)

A web, tal como é organizada, não consegue atender várias demandas informacionais além de possuir problemas como:

- **Falta de capacidade em responder questões práticas ou específicas**. Mika (2007) chama atenção para o fato de que mesmo se a informação estivesse disponível e fosse evidente para o usuário, o computador poderia ficar limitado à visão de apenas um punhado de caracteres que mesmo comparando a palavras chave fornecidas pelo usuário, não haveria qualquer entendimento do significado delas. Na maioria dos casos, no entanto, a falta de conhecimentos é devido à falta de algum tipo de conhecimento prévio que apenas o humano possui. Como vários conhecimentos básicos são completamente ausentes do contexto da página web, os computadores não tem como responder corretamente às demandas. Segundo Stuckenschmidt (2012), não é simplesmente encontrar um documento ou uma página da web, mas encontrar uma informação específica, por exemplo, coletar informações sobre uma única pessoa, ou obter resposta para questões como: “O que faz o clima mudar?”.

- **Falta de padrões**. Bauer e Kaltenböck (2012) e Ding, Peristeras e Hausenblas (2012) destacam a falta de padrões na web. Shah (2013), exemplifica situações em quatro contextos diferentes: (1º) para um indivíduo: seria mais simples para uma pessoa ligar, ouvir e interagir com todas as suas redes ao invés de abrir múltiplos fluxos isolados de forma separada; (2º) para uma organização: vários fornecedores diferentes poderiam colocar informações num único fluxo de dados para facilitar a leitura pela organização; (3º) para o cliente: se o serviço de atendimento ao cliente (SAC) não for padronizado é bem possível que ele recorra para as redes sociais para obter um atendimento mais adequado; (4º) entre organizações: considerando que parcerias ou compromissos entre empresas envolvem trabalhadores, então a padronização nas formas de se comunicar facilitaria o processo.
- **Falta de capacidade para processamento via máquina**. É necessário disponibilizar informações de tal forma que os computadores possam processar, extrair ou adicionar conhecimentos, para auxiliar a responder demandas diversas (BERNERS-LEE *et al.*, 2001; MIKA, 2007; AUER *et al.*, 2013).
- **Alto custo na publicação em ambientes distribuídos**. Para Bauer e Kaltenböck (2012), a ideia básica da Web Semântica é fornecer formas eficientes quanto ao custo para publicar informação em ambientes distribuídos. Para os custos o estabelecimento de padrões é crucial entre os sistemas que transmitem e recebem a informação, pois ambos devem convertê-la ou mapeá-la em três níveis: sintaxe, esquema e vocabulário.

A Web Semântica trouxe possiilidades para tentar solucionar esses quatro problemas citados. Ela é a aplicação de tecnologias avançadas do conhecimento, a fim de preencher a lacuna de conhecimento entre homem e máquina e tem o World Wide Web Consortium (W3C) como sua principal força propulsora (MIKA, 2007), particularmente com o projeto *Linking Open Data*[12], um esforço da comunidade de base fundada em 2007 (HEATH; BIZER, 2011).

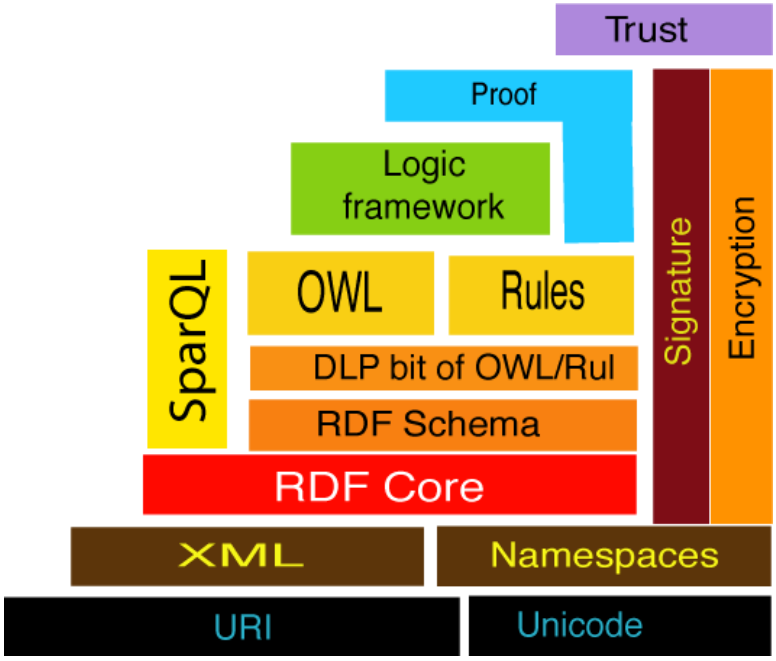
Berners-Lee (2005) sugeriu uma arquitetura para Web Semântica, mostrada na Figura 5, para organizar de forma simples e em camadas as várias linguagens e esquemas usados no seu contexto. Dessa forma, cada camada da figura representa um nível de complexidade, do mais simples, na base, até o mais complexo, no topo. Sendo que, uma determinada camada utiliza-se dos elementos das camandas inferiores. As camadas da Figura 5 são aqui abordadas

[12] Projeto *Linking Open Data*: <http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>.

de forma resumida e macro conceitual, sem ainda tratar especificamente das tecnologias citadas, pois estas são explicadas mais adiante ao longo dessa seção.

- **URI/UNICODE**: é a camada que faz a ligação direta da web com a Web Semântica, utilizando-se, sempre que possível, de um padrão internacional, como o Unicode;
- **XML/Namespaces**: integra a web com a Web Semântica aproveitando as informações e documentos já existentes e generalizando a linguagem de HTML para XML;
- **RDF Core**: é uma camada que foi especificamente desenvolvida para Web Semântica e representa uma rede de informações formada por triplas, onde cada uma é composta por dois nós conceituais e um predicado que faz uma ligação descritiva entre eles;
- **RDF Schema**: acrescenta propriedade taxonômica na camada anterior para determinar de forma melhor o relacionamento entre as informações e também facilitar o seu acesso;
- **DLP bit of OWL/Rul**: por intermédio de linguagens de ontologia, essa camada formaliza, padroniza e melhora a expressividade da semântica das informações e, consequentemente, a interoperabilidade e o seu acesso.
- **OWL/Rules**: essa camanda trabalha com a informação em um nível mais elevado onde regras são usadas para representação de conhecimento;
- **Logic framework**: oferece motores de inferência que trabalham com lógica de primeira ordem para manipulação do conhecimento;
- **Proof**: essa camada tem a capacidade de mostrar como que se chegou a um resultado com o uso de uma máquina de inferência de lógica de primeira ordem;
- **Trust**: finalmente, a camada de mais alto nível, representa a confiança da informação e conhecimento fornecidos pela Web Semântica, usando ferramentas tais como chaves públicas, assinatura digital e criptografia, mostradas na figura como ‘Signature’ e ‘Encryption’.

Figura 5 – Arquitetura da Web Semântica



Fonte: Berners-Lee (2005)

A Web Semântica tem vários desafios a serem vencidos. Por exemplo, Stuckenschmidt (2012) sinaliza a escalabilidade como o desafio mais significativo da gerência de dados semânticos da web. Stuckenschmidt também expõe o conhecido gargalo na aquisição de conhecimento (*Knowledge Acquisition Bottleneck*) como um problema associado aos modelos semânticos. Segundo o autor, a criação de modelos semânticos de forma manual é uma opção cara e inviável, por outro lado a criação de modelos semiautomáticos é complexa e necessita de vários métodos e combinação de várias abordagens, tal como *bootstraping*[13]. Apesar de Paulhem (2013) relatar que a abordagem de enriquecimento automática deva ser sempre buscada, Stuckenschmidt indica a participação do usuário de forma ativa no processo de diminuição do gargalo de conhecimento, sinalizando o *crowdsourcing*[14] como uma boa opção. Howe (2006) exemplifica vários casos de sucesso de *crowdsourcing*. Observa-se na web 2.0 que os usuários estão dispostos a fornecer conteúdo bem como metadados (MIKA, 2007). Muitas das evoluções que tanto fervilham nas tecnologias da web e Web Semântica poderiam acontecer com auxílio do *crowdsourcing* envolvendo pessoas nesse processo colaborativo e assim alavancar muitos projetos de desenvolvimento de novas tecnologias.

[13] *Bootstraping* é um processo de retroalimentação onde dados de saída são usados também como entrada.
[14] *Crowdsoursing* é uma prática que se utiliza da inteligência coletiva, geralmente de voluntários espalhados pela web, para resolver problemas, criar conteúdo e soluções, desenvolver novas tecnologias entre outros.

Como a Web de Dados é baseada em padrões e um modelo de dados comum, torna-se possível implementar aplicações genéricas que operam sobre o espaço de dados completo da web (HEATH; BIZER, 2011). Desde o início do século atual, Berners-Lee *et al.* (2001) sinalizaram que a web de dados, referida como Web Semântica em seu artigo, poderia colaborar com a evolução do conhecimento humano como um todo, pois ela não é simplesmente uma ferramenta para realização de tarefas individuais, pois na medida em que o conhecimento humano é construído de forma colaborativa e a Web Semântica é munida de ontologias e agentes de software, isso poderá auxiliar a comunicação e o trabalho colaborativo entre diversos povos de culturas e/ou línguas diferentes. Esse é um processo lento e contínuo, contudo, muitos resultados interessantes já existem nos dias de hoje.

2.3.2 Dados abertos

Dados abertos são conteúdos que podem ser livremente usados, modificados e compartilhados por qualquer um e para qualquer finalidade, sujeitos, no máximo, a medidas que preservem a sua proveniência e abertura (OPEN DEFINITION, 2015). Na página web da Open Definition existe a descrição detalhada dos princípios que definem como deve ser a abertura sobre dados e conteúdos. Esses princípios são organizados em dois grupos: (1º) a abertura do trabalho, relativa a quatro requisitos (disponibilidade, acessibilidade, capacidade de ser lida por máquina, formato aberto); (2º) sobre a licença de abertura, relativa aos requisitos de permissão com nove exigências (uso, redistribuição, modificação, separação, compilação, não discriminação, propagação, propósito, sem custo) e relativa às condições de aceitabilidade com mais sete exigências (atribuição, integridade, compartilhamento nos mesmos termos, aviso do tipo de licença, restrição técnica, não agressão).

Segundo o manual Open Data Handbook (2012) existem três regras-chave para realizar o processo de abertura dos dados de uma organização:

(i) Mantenha simplicidade: a abertura dos dados deve ser feita de forma gradual;

(ii) Envolva-se com a comunidade: grande parte dos dados abertos vão ainda passar por mediadores que irão transformá-los para uma aparência mais acessível aos usuários finais, portanto, se envolver com todos esses usuários finais ou intermediários é fundamental para que os dados selecionados para serem abertos sejam relevantes; e

(iii) Cuide dos receios e mal-entendidos: muitos temores irão surgir principalmente em grandes instituições, como os governos, e devem ser identificados e trabalhados o quanto antes.

Além disso, esse manual indica quatro passos principais para realizar a abertura dos dados, que de forma resumida são:

(i) A escolha do conjunto dos dados para abrir;

(ii) A escolha de uma licença livre compatível com os direitos de propriedade intelectual dos dados;

(iii) A disponibilização dos dados em um grande volume e num formato útil; e

(iv) A catalogação necessária para que os dados sejam facilmente localizáveis.

A divulgação dos dados abertos para a sociedade é uma etapa importante e deve ser feita de forma eficiente. Vários países, tais como Brasil[15], Estados Unidos[16] e Inglaterra[17], fazem a divulgação através de sites que centralizam e facilitam o acesso aos dados abertos de seu governo oferecendo serviços, dicas, manuais, guias, interação com os usuários e várias outras ações e elementos associados ao tema. Um bom exemplo de aplicação de dados abertos no Brasil é o Portal da Transparência do Governo Federal[18] que, funcionando desde 2004, tenta assegurar a boa e correta aplicação dos recursos públicos aumentando a transparência da gestão pública e permitindo que o cidadão acompanhe como o dinheiro público está sendo utilizado ajudando a fiscalizá-lo. Com o argumento de que a transparência é o melhor antídoto contra corrupção, esses dados abertos induzem os gestores públicos a agirem com responsabilidade e permitem que a sociedade, munida de informações, colabore com o controle das ações de seus governantes. Apesar da boa proposta e esforço realizado pelo Portal da Transparência, existem alguns pontos de melhora tal como detectado na pesquisa de Nazario, Silva e Rover (2012), que fizeram uma análise da qualidade da informação detectando três pontos negativos: (i) algumas informações estão implícitas, ou seja, só são reveladas após o tratamento num software como uma planilha eletrônica; (ii) as informações estão distribuídas em diferentes consultas não permitindo o cruzamento de dados; (iii) falta detalhamento para leigos pois parece que as informações são fornecidas apenas para o entendimento de especialistas.

Outra variável importante que determina a qualidade do acesso à informação é a definição de formatos de abertura dos dados que, apesar de todo o esforço legislativo no Brasil, desde a Constituição, (BRASIL, 1998), a Lei de Acesso a Informação (BRASIL, 2011)

[15] Portal de dados abertos do governo do Brasil: <http://dados.gov.br/>.
[16] Portal de dados abertos do governo dos Estados Unidos: <http://dados.gov.br/>.
[17] Portal de dados abertos do governo da Inglaterra: <http://data.gov.uk/>.
[18] Portal da transparência do governo federal: <http://www.portaltransparencia.gov.br/>.

e a sua regulamentação (BRASIL, 2012), ainda é um dos grandes problemas. Segundo o Manual dos Dados Abertos para Desenvolvedores (2011), a definição do formato em que a informação é aberta deveria ser disponibilizada pelo governo, pois a "[...] lei diz que os dados devem ser expostos à população, mas não especifica como [...]". Por isso, ainda segundo o manual, "[...] muitas vezes os esforços de transparência, tão valorizados pela administração pública de nosso país, acabam sem a repercussão ou resultados desejados". Portanto, é importante destacar que apenas disponibilizar em formato aberto não é suficiente, pois, na prática, é necessário que eles sejam expostos em formato compreensível por máquina e também por humanos.

O Manual dos Dados Abertos para Desenvolvedores (2011) exemplifica várias experiências de sucesso, realizadas com dados do governo brasileiro, como o projeto Legisdados que extrai dados não estruturados do Legislativo Brasileiro e os disponibiliza em formatos compreensíveis por máquina. Isso facilita o trabalho de desenvolvedores que querem criar sistemas integrados com essas informações. O manual citado também alerta que, se for realizado "[...] um trabalho mais elaborado e organizado para disponibilizar dados, haverá benefícios para todos: o governo receberá sugestões e haverá uma participação mais eficiente da sociedade, a qual saberá mensurar melhor o trabalho das autoridades". No futuro, os mercados de dados emergentes serão os mecanismos que transformarão as contribuições voluntárias para agregação de valor em um setor de negócios rentável (DING; PERISTERAS; HAUSENBLAS, 2012).

Por outro lado, problemas também foram detectados na aplicação da Lei de Acesso à Informação. Pedroso, Tanaka e Cappelli (2013) indicam falhas nos dados abertos já disponibilizados pelo governo, alertando que "são gritantes as inadequações de formatos, que impedem a interoperabilidade, assim como é evidente a existência de dados incompletos, desatualizados, incompreensíveis por máquina e dependentes de licenças proprietárias [...]". Os autores ainda destacam que "[...] tais deficiências refletem a ausência de processos organizacionais e da adoção de padrões existentes, assim como a falta de pessoas capacitadas a darem conta das atividades necessárias ao planejamento e implantação dos processos [...]". Finalmente eles observam que muitos problemas que as organizações estão enfrentando nesse contexto ainda não tem solução sistematizada e ainda indicam a necessidade de muitos estudos e pesquisa na área.

2.3.3 Dados ligados e dados abertos ligados

O termo dados ligados, conhecido internacionalmente por *Linked Data*[19], refere-se ao uso da web para criar links entre dados de fontes diferentes. Essas fontes podem ser bancos de dados mantidos por duas organizações em diferentes localizações geográficas, ou simplesmente sistemas heterogêneos dentro de uma organização que normalmente não seriam ligados. Tecnicamente, dados ligados referem-se aos dados publicados na web, de tal forma que sejam legíveis por máquina, com significado explicitamente definido, e ligado a outros conjuntos de dados externos. Segundo Health e Bizer (2011), dados ligados dizem respeito a um conjunto de melhores práticas para a publicação, compartilhamento e ligação de dados, informações e conhecimento sobre a web. Assim como hiperlinks na web clássica conectam documentos em um único espaço de informação global, dados ligados usam hiperlinks para conectar dados diferentes em um único espaço de dados global.

No contexto conceitual dos dados ligados, existem elementos que são fundamentais para o seu entendimento:

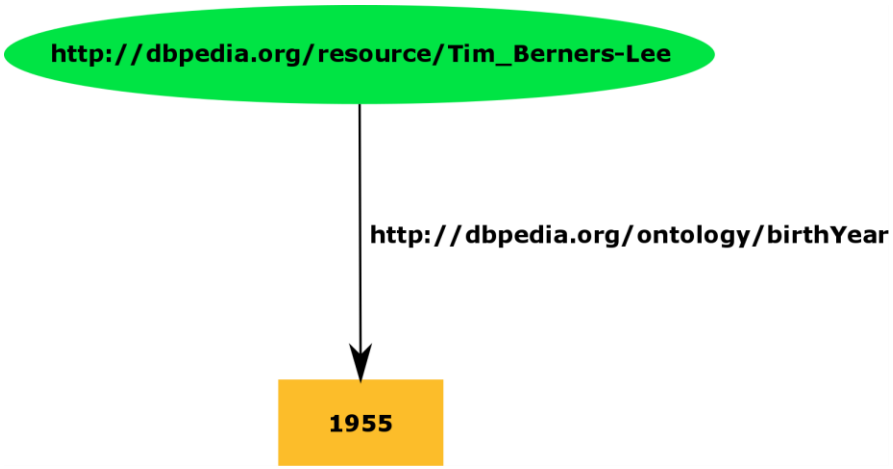
- *Resource Description Framework* (RDF): é uma estrutura para representar informações na web e, principalmente, permitir a interoperabilidade de metadados. Ela é uma tripla composta por um sujeito/recurso (*subject*), um predicado/propriedade (*predicate*) e um objeto/valor (*object*), onde o predicado tipifica o relacionamento existente entre o sujeito e o objeto. A Figura 6 representa o relacionamento ‘Tim Berners-Lee nasceu em 1955’, na forma de uma tripla RDF. Nesse grafo, a elipse nomeada representa o sujeito/recurso, a seta nomeada representa o predicado/propriedade e o quadrilátero nomeado representa o objeto/valor. Assim, o recurso ‘http://dbpedia.org/resource/Tim_Berners-Lee’, está associado ao valor ‘1995’ por intermédio da propriedade ‘http://dbpedia.org/ontology/birthYear’. A sintaxe detalhada do RDF pode ser obtida no site da W3C[20].
- *Uniform Resource Identifier*[21] (URI), representa um recurso na tripla RDF. Pode ser classificado como um nome, *Uniform Resource Name* (URN), ou seja, a identidade de um item, ou como um localizador *Uniform Resource Locator* (URL) que é um endereço web do recurso. Um URL é popularmente conhecido por endereço de internet. A Figura 6 apresenta dois casos de localização única: o recurso

[19] *Linked Data* é conceituado em: <http://linkeddata.org/>.
[20] Sintaxe do RDF: <https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/>.
[21] URI é especificado em: <http://tools.ietf.org/html/rfc3305>.

‘http://dbpedia.org/resource/Tim_Berners-Lee’ e a propriedade ‘http://dbpedia.org/ontology/birthYear’, que apontam para páginas na web que descrevem o recurso e a propriedade, respectivamente. URI não é só uma tecnologia, mas também é considerada a chave para a universalidade da web (BERNERS-LEE, 2010).

- *Internationalized Resource Identifier*[22] (IRI): também representa um recurso na tripla RDF, porém, é uma generalização ou internacionalização do URI, já que este último é limitado a um subconjunto do conjunto de caracteres ASCII[23], e o IRI pode conter caracteres de conjuntos muito maiores tal como o Unicode[24].

Figura 6 – Exemplo de uma tripla RDF

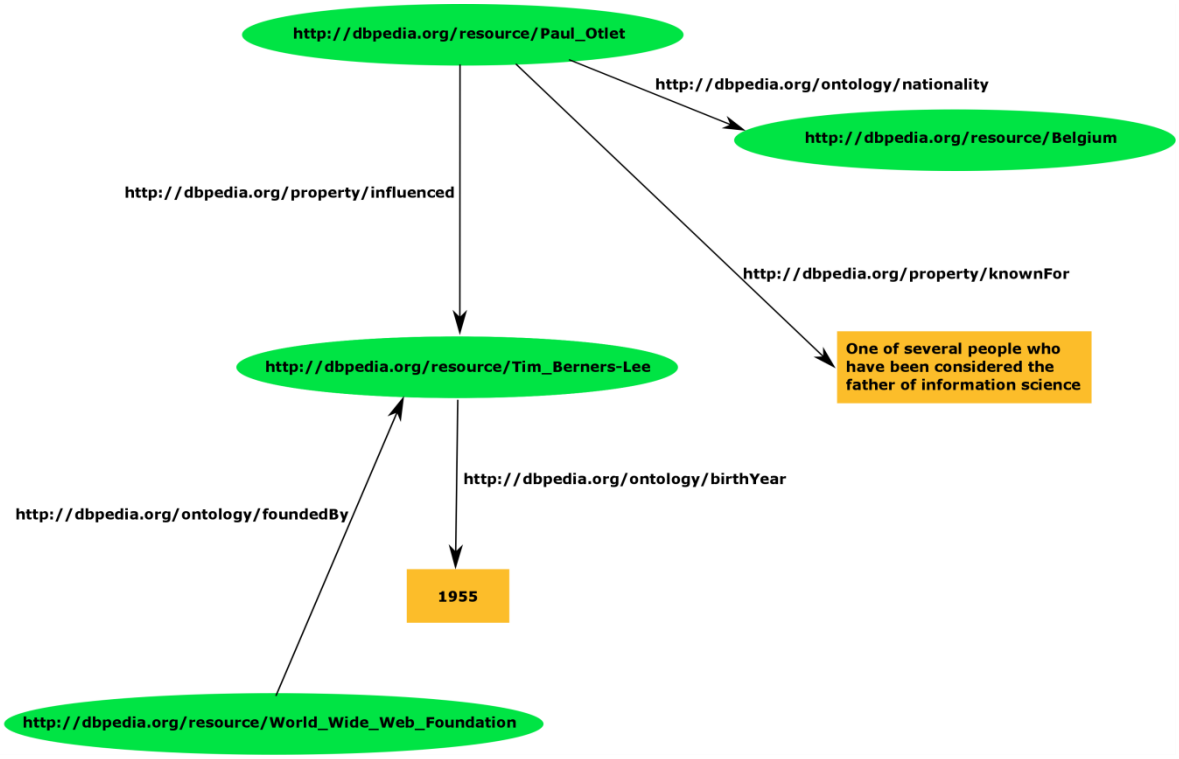


Fonte: Elaboração própria

Um conjunto de triplas RDF é chamado de grafo RDF, e pode ser visualizado como um conjunto de vértices (nós) e arestas (links), onde cada tripla determina a presença de dois vértices e uma aresta no grafo. Do ponto de vista informacional, uma tripla RDF é formada por um recurso ligado, por uma propriedade, a um valor. Sendo que o valor pode denotar qualquer coisa, incluindo coisas físicas, documentos, conceitos abstratos, números, strings ou outros recursos. A Figura 7 representa um grafo real extraído da base de dados ligados DBpedia[25].

[22] IRI é especificado em: <http://tools.ietf.org/html/rfc3987>.
[23] ASCII, sigla de American Standard Code for Information Interchange, é uma tabela básica de caracteres: <http://www.asciitable.com/>.
[24] Unicode é uma tabela ampla de caracteres internacionais: <http://unicode-table.com/en/>
[25] DBpedia: é uma base de conhecimento que segue os preceitos dos dados abertos ligados. Ela será abordada em detalhes seção 2.3.5.

Figura 7 – Exemplo de uma rede de triplas RDF extraída da DBpedia



Fonte: Elaboração própria

A rede ou grafo da Figura 7 resulta da expansão, com mais quatro RDFs, a tripla da tripla mostrada na Figura 6, formando um grafo RDF ou uma rede informacional onde observam-se as seguintes relações: ‘Paul Otlet tem nacionalidade belga’, ‘Paul Otlet é conhecido por ser uma das várias pessoas consideradas como pai da Ciência da Informação’, ‘Paul Otlet influenciou Tim Berners-Lee’, ‘Tim Berners-Lee nasceu em 1955’ e ‘A World Wide Web Foundation foi fundada por Tim Berners-Lee’. Há várias formas de extrair informações de uma base de dados ligados. Por exemplo, as informações da Figura 7 foram extraídas da DBpedia por intermédio de consultas escritas na linguagem SPARQL[26] e utilizando-se o terminal SNORQL[27].

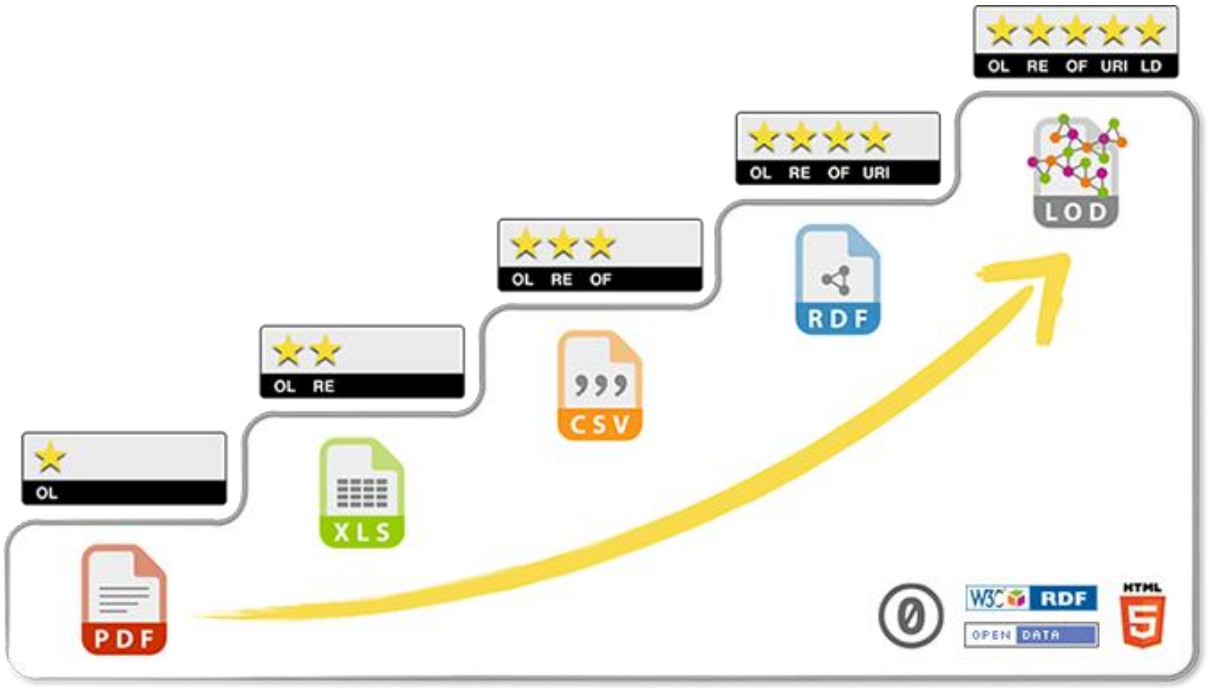
O termo ‘Dados Abertos Ligados’, conhecido internacionalmente por ‘*Linked Open Data’* (LOD), refere-se a dados ligados num contexto de dados abertos (seção 2.3.2). Muito conhecida e referenciada na literatura a classificação das 5 estrelas idealizada por Berners-Lee (2006), atualizada pelo próprio autor em 2010 e representada pela Figura 8, serve para ranquear dados publicados na web, onde cada nível acumula as características no nível

[26] SPARQL é uma linguagem de consulta em LOD, derivada da linguagem SQL de consulta a bancos de dados. Especificação disponível em <https://www.w3.org/TR/sparql11-query/>.
[27] SNORQL: é um terminal para acesso aos dados da DBpedia por intermédio de consultas SPARQL. Disponível em <http://dbpedia.org/snorql/>.

anterior e apresenta novas características, começando pelo mais simples (uma estrela) até o mais completo (cinco estrelas), que representa os dados abertos ligados:

- **Uma estrela**: representa a possibilidade dos dados serem lidos na web, em qualquer formato, porém, sob uma licença aberta – exemplo, um arquivo do tipo PDF;
- **Duas estrelas**: os dados são em um formato estruturado e podem ser lidos por uma máquina – exemplo, um arquivo XLS de uma planilha eletrônica proprietária como o Excel[28];
- **Três estrelas**: os dados não usam formato proprietário – exemplo, os dados de um arquivo de planilha eletrônica gravados num formato livre CSV[29];
- **Quatro estrelas**: os dados usam URI para a identificação e, assim, permitem que outras pessoas possam publicar material apontando para eles;
- **Cinco estrelas**: os dados são efetivamente ligados com os dados de outras fontes, ou seja, se comportam plenamente como dados abertos ligados ou *linked open data* (LOD).

Figura 8 – Classificação dos dados ligados



Fonte: http://5stardata.info/

[28] Copyright Microsoft Corporation: <https://www.microsoft.com>.
[29] Formato CSV: valores separados por vírgula (*comma-separated values*).

A atualização dessa classificação que Tim Berners-Lee fez em 2010 ocorreu para o item referente a uma estrela, acrescentando a exigência de que os dados fossem abertos. Como o padrão cinco estrelas herda as características do padrão uma estrela, então os dados deverão ser abertos em todo os casos. Hausenblas e Kim (2015) fizeram uma interessante interpretação sobre a classificação de Tim Berners-Lee acrescentando o custo benefício para cada um dos níveis, tanto do lado do consumidor da informação quanto do lado de quem publica a informação na web. De uma maneira geral o custo de tempo aumenta, para quem publica, na medida em que se aumenta a quantidade de estrelas. Há um ganho muito significativo em possibilidades, para o consumidor da informação, com esse mesmo aumento do nível de estrelas.

Health e Bizer (2011), Auer *et al.* (2013) destacam benefícios para o uso de dados ligados sintetizados aqui em seis tópicos: (i) *Uniformity*: os dados publicados compartilham um único modelo de dados estruturados, chamado RDF; (ii) *De-referencability*: além das URIs serem usadas para a identificação de entidades, elas também podem ser usadas, tal como URLs, para a localização e recuperação de recursos descritos e representados por entidades na internet. A característica *de-referencability* é reconhecida por Cyganiak *et al.* (2014) como sendo a melhor contribuição das IRIs e fundamental para o conceito de dados ligados; (iii) *Coherence*: quando uma tripla RDF contém URIs de namespaces diferentes então é estabelecida uma ligação entre a entidade identificada pelo sujeito com a entidade identificada pelo objeto. Itens de dados são ligados através do tipo da RDF; (iv) *Integrability*: considerando que todas as fontes de dados referenciados compartilham o modelo de dados RDF, que é baseado em um único mecanismo para representar as informações, fica fácil alcançar uma integração semântica sintática simples de diferentes conjuntos de dados ligados; (v) *Timeliness*: a publicação e atualização de dados ligados é simples, não precisando mais de gasto de tempo para as etapas de extração, transformação e carregamento; e (vi) *Self-descriptive data*: dados ligados facilitam a integração de dados de diferentes fontes, baseando-se em vocabulários compartilhados, tornando as definições destes vocabulários recuperáveis, e ao permitir condições de diferentes vocabulários para ser ligados uns aos outros por ligações de vocabulário.

Berners-Lee (2006) elaborou quatro princípios fundamentais para a criação de dados ligados: (1) use URIs para dar nomes as coisas; (2) use HTTP URIs para que esses nomes possam ser acessados; (3) quando um URI é acessado, responda com dados úteis, utilizando padrões da web tais como RDF e SPARQL; e (4) inclua links para outras URIs para facilitar a busca por novos dados.

Ainda sobre a produção de dados ligados, Auer *et al.* (2013) descrevem o seu ciclo de vida por intermédio de etapas que se completam mutualmente e interagem uma com as outras: (i) *Extraction*: informações representadas em forma não estruturada ou que obedecem a outros formalismos de representação estruturados ou semiestruturados devem ser mapeados para o modelo de dados RDF; (ii) *Storage/Quering*: considerando um conjunto de dados no formato RDF, mecanismos armazenam, indexam e consultam esses dados de forma eficiente; (iii) *Authoring*: usuários devem ter a oportunidade de criar novas informações estruturadas ou corrigir e ampliar as já existentes; (iv) *Linking*: se diferentes dados publicados fornecem informações sobre as mesmas entidades relacionadas, então devem ser realizadas ligações entre esses elementos; (v) *Enrichment*: considerando que os dados ligados possuem instâncias de dados, é possível observar falta de classificação, estrutura e esquema de informações, e essa deficiência pode ser resolvida através de abordagens que enriqueçam os dados com estruturas de nível mais elevado, a fim de permitir a agregação e consulta aos dados de modo mais eficiente; (vi) *Quality Analysis*: tal como acontece com a web de documentos, a Web de Dados contém uma variedade de informações de diferentes qualidades, por isso, é importante elaborar estratégias para avaliar a qualidade dos dados publicados na Web de Dados; (vii) *Evolution & Repair*: se problemas são detectados, então é necessário empregar estratégias para repará-los e para apoiar a evolução dos dados vinculados; e (viii) *Search, Browsing & Exploration*: usuários devem possuir poderes para navegar, pesquisar e explorar as informações de estrutura disponível na Web de Dados de forma rápida e amigável.

O projeto *Linking Open Data*[30] é o exemplo mais visível da adoção e aplicação dos princípios de LOD (BIZER; HEATH; BERNERS-LEE, 2009). Nesse projeto existe uma boa quantidade de entidades disponíveis em um volume estimado de 50 bilhões de elementos oriundos de muitos domínios diferentes, como geografia, meios de comunicação, biologia, química, economia, energia, etc. (BAUER; KALTENBÖCK, 2012). O *Linking Open Data* tem como objetivo estender a web como um bem comum de dados, através da publicação de vários conjuntos de dados abertos como RDF na web, e estabelecendo ligações entre os itens de dados RDF a partir de diferentes fontes de dados. O projeto apoio do movimento Open Data, que visa tornar os dados disponíveis gratuitamente para todos. Já existem vários conjuntos de dados interessantes abertos disponíveis na Web, como por exemplo:

[30] Projeto *Linking Open Data*, disponivel em: <http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>.

Wikipedia[31], Wikibooks[32], Geonames[33], MusicBrainz[34], WordNet[35], a bibliografia DBLP[36] etc.

Apesar do sucesso premente dos dados ligados, existem problemas e reflexões importantes detectados por alguns autores. Berners-Lee (2010) alerta que várias "ameaças" têm surgido para a formação de ilhas na web, e um dos fatores para esta possibilidade de isolamento acontece porque cada pedaço de informação não possui um URI. Zaveri *et al.* (2015), Képéklian, Curé e Bihanic (2015) sinalizam para uma preocupação mais recente com a qualidade nos dados ligados. Berners-Lee e O'Hara (2013), Képéklian, Curé e Bihanic (2015) sinalizam para a questão da privacidade dos dados, pois numa situação, por exemplo, de integração com redes sociais, informações pessoais e privadas estariam abertas. Sobre a necessidade de crescimento das bases de dados ligados, Stuckenschmidt, Noessner e Fallahi (2012) sugerem que o próprio usuário trabalhe no enriquecimento da sua base de dados. Segundo esses autores, a abordagem centrada no usuário é mais vantajosa em comparação com aquela em que as tarefas de integração de dados são realizadas por profissionais de tecnologia da informação. Essa abordagem centrada no usuário baseia-se em um modelo cognitivo que permite que pessoas com pouco ou nenhum conhecimento possam integrar seus dados. Paulheim (2013) alerta para a variedade dos dados abertos ligados, pois apesar deles estarem sendo construídos em padrões bem definidos, há maneiras diferentes para fornecê-los. Nessa mesma linha, Auer *et al.* (2013) sinalizam que o desenvolvimento de abordagens de pesquisa, padrões, tecnologias e ferramentas para apoiar o ciclo de vida dos dados ligados é um dos principais desafios atuais, e podem causar um impacto substancial sobre a ciência, a economia, a cultura e a sociedade em geral.

---

[31] Wikipedia é um projeto de enciclopédia multilíngue de licença livre, baseado na web, escrito de maneira colaborativa e administrado pela Fundação Wikimedia. Disponível em: <http://www.wikipedia.org/>.

[32] Wikibooks é um projeto da Wikimedia Foundation dedicado ao desenvolvimento colaborativo de textos didáticos de conteúdo livre. Disponível em: <http://www.wikibooks.org/>.

[33] GeoNames é uma base de dados geográfica livre com informações sobre mais de 8 milhões de lugares. Disponível em: <http://www.geonames.org/>.

[34] MusicBrainz é uma enciclopédia de musica aberta que armazena metadados de músicas e os disponibiliza ao público. Disponivel em: <http://musicbrainz.org/>.

[35] WordNet é um grande banco de dados léxico de Inglês mantido pela National Science Foundation. Disponível em: <http://wordnet.princeton.edu/>.

[36] DBLBibliografy é um banco de dados com bibliografia da Ciência da Computação, mantido pelo DBLP Team. Disponível em: <http://dblp.uni-trier.de/db/>.

2.3.4 Ontologias para dados ligados

A estruturação do conhecimento pode seguir níveis diferentes de sofisticação, com o emprego de sistemas de organização do conhecimento. Segundo Képéklian, Curé e Bihanic (2015), quatro sistemas conhecidos, do mais simples ao mais sofisticado, são:

(i) **Vocabulário controlado**: é conjunto de termos (e eventualmente suas definições) organizados para redução da ambiguidade por meio de referências cruzadas entre termos, especialmente homógrafos, sinônimos e polissêmicos, indicando os termos autorizados para uso em atividades, como de indexação (HEDDEN, 2008). Tal como um glossário, o vocabulário controlado tem como objetivo estabelecer o conjunto de termos que formam um vocabulário inibindo disparidades, problemas de sinônimos e grafia errada (KÉPÉKLIAN; CURÉ; BIHANIC, 2015) e são, em essência, um acordo sobre o significado de um conjunto de termos (MIKA, 2007).

(ii) **Taxonomia**: é um conjunto de termos hierarquicamente organizados (HEDDEN, 2008) permitindo a existência de subconjuntos em um vocabulário controlado (KÉPÉKLIAN; CURÉ; BIHANIC, 2015);

(iii) **Thesaurus:** conjunto de termos organizados com referências cruzadas, que permite representação de relações hierárquicas, associativas e de equivalência, além de incluir notas sobre utilização dos termos (HEDDEN, 2008). Dessa forma, ele vai além da possibilidade de subgrupos da taxonomia, por permitir o uso de outras propriedades para descrever as relações entre os termos (KÉPÉKLIAN; CURÉ; BIHANIC, 2015); e

(iv) **Ontologia:** sistema de organização do conhecimento que oferece suporte para representar um número ilimitado de tipos de relacionamentos e atribuição de significado único pra conceitos (KÉPÉKLIAN; CURÉ; BIHANIC, 2015).

No contexto da Ciência da Informação e da Ciência da Computação uma ontologia define um conjunto de primitivas de representação, compostas por classes (ou conjuntos) e atributos (ou propriedades) com os quais modela um domínio do conhecimento (GRUBER, 2009). O autor ainda explica que, num contexto de banco de dados, as ontologias ficam em um nível semântico, enquanto que o próprio esquema de banco de dados representa modelos de dados no nível lógico ou físico. Esse nível semântico permite independência, e por isso as ontologias são usadas para integrar bancos de dados heterogêneos, permitindo a

interoperabilidade entre sistemas diferentes e a especificação de interfaces baseadas em conhecimento, para serviços independentes.

O uso de ontologias é um importante elemento na concepção da Web Semântica, pois elas definem a natureza das informações, sua classificação e as relações entre termos além de oferecer precisão nas pesquisas, evitando a ambiguidade de palavras chaves (BERNERS-LEE *et al.*, 2001). A maioria das bases de dados abertos ligados vem com alguma semântica explícita por intermédio de vocabulários que contém informações semânticas na forma de declarações de ontologias que oferecem informações valiosas para o seu enriquecimento (PAULHEIM, 2013). Ontologias fazem parte dos padrões W3C para a Web Semântica, sendo usadas para especificar vocabulários conceituais padrão, para a troca de dados entre sistemas, fornecimento de serviços para responder a perguntas, publicação de bases de conhecimento reutilizáveis e oferta de serviços, facilitando a interoperabilidade entre múltiplos sistemas e bancos de dados heterogêneos (GRUBER, 2009). Para satisfazer a diferentes necessidades, o W3C (W3C - ONTOLOGIES, 2015) oferece uma ampla variedade de técnicas para descrever e definir diferentes formas de vocabulários como formato padrão: RDF e *RDF Schemas*, *Simple Knowledge Organization System* (SKOS), *Web Ontology Language* (OWL), e *Rule Interchange Format* (RIF). A escolha entre essas diferentes formas é dependente da complexidade e rigor necessários na aplicação. Um exemplo de aplicação de ontologia para dados abertos ligados é brevemente descrito na próxima subseção, 2.3.5, no contexto da DBpedia.

2.3.5 DBpedia

A DBpedia é uma base de conhecimento de dados abertos ligados que segue a classificação das 5 estrelas ditadas por Berners-Lee (2006) e discutida na subseção 2.3.3. Ela é um esforço comunitário para extrair informações estruturadas da Wikipedia[37] e tornar essas informações disponíveis na web permitindo sofisticadas consultas (AUER *et al.*, 2007) à sua base de conhecimento, além de cobrir uma grande quantidade de áreas, sendo amplamente usada pela comunidade de pesquisa e por diversas aplicações (LEHMANN *et al.*, 2015).

A Wikipedia, que fornece as informações da DBpedia, surgiu em 2001 e desde então cresceu exponencialmente em quantidade de artigos, sendo hoje o sexto site mais consultado

[37] Wikipedia é uma enciclopédia livre, colaborativa, baseada na web e mantida pela Wikimedia Foundation, Inc. Disponível em <https://www.wikipedia.org/>.

em toda web, segundo ranking do portal Alexa[38]. Ela se transformou num alicerce da cultura, com uma rapidez que não era esperada, talvez por conta de seu relacionamento sinergético e não planejado com o buscador Google (GLEICK, 2011). Segundo a própria Wikipedia, em 2015 ela alcançou 38 milhões de artigos em mais de 250 línguas. Ela é um dos melhores exemplos de criação de conteúdo colaborativo (LEHMANN *et al.*, 2015). Devido à sua vulnerabilidade quanto ao vandalismo ou mesmo contribuições menos especialistas, houve um tempo, principalmente no início de sua trajetória, que ela não era recomendada por acadêmicos, jornalistas etc. Mais recentemente a Wikipedia é sustentada por uma "conspiração gigantesca" entre programas de computador e comunidades de humanos voluntários, para garantir que seus artigos sejam condizentes com a realidade (GLEICK, 2011).

Segundo a página de estatísticas da DBpedia[39], em outubro de 2015 existiam 6,2 milhões de entidades da web semântica nela, sendo que 5 milhões foram devidamente classificadas usando-se ontologias consistentes. A DBpedia usa o padrão RDF para representar informações e possui 8,8 bilhões de triplas RDF. A Figura 9 representa a arquitetura da DBpedia. A infraestrutura do seu servidor (*Virtuoso Universal Server*) oferece aos usuários acesso aos seus dados RDF por intermédio de três canais: (1) *HTML Consumers*: acesso simples via páginas HTML, (2) *RDF Consumers*: acesso direto aos RDFs pela web, (3) *SPARQL Clients*: usando a linguagem de consulta SPARQL e um terminal de consulta (*SPARQL Endpoint*), tal como o SNORQL.

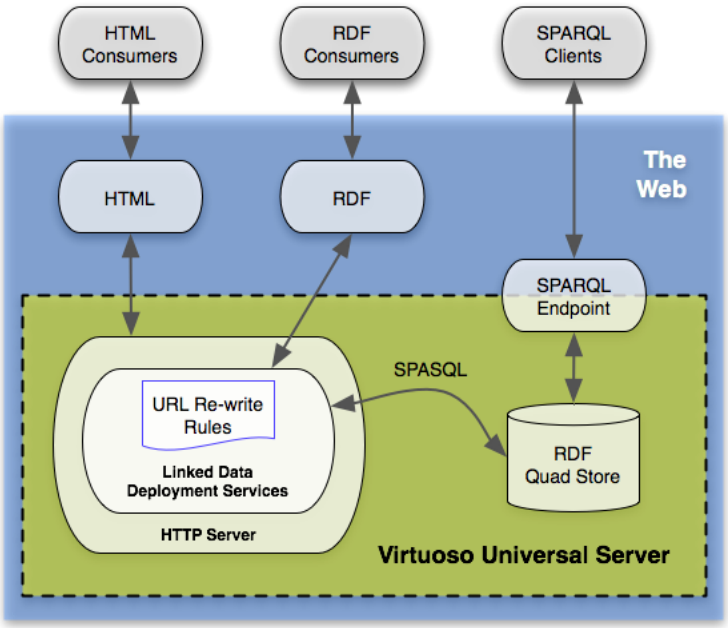
Ainda segundo a página de estatísticas da DBpedia, a ontologia usada, que é atualizada pela própria comunidade de usuários, possui 739 classes que formam uma hierarquia 2.695 propriedades obedecendo um limite máximo de níveis para que seja mantida boa visibilidade e navegabilidade. A Figura 7, da seção 2.3.3, que apresentou um grafo RDF extraído da DBpedia, possui exemplos de elementos da ontologia da DBpedia que foram usados em relações com os recursos: *influenced*, *foundedBy*, *bithYear*, *nationality*, *knownFor*.

O projeto DBpedia está em constante evolução e possui desafios a serem vencidos. Um deles é saber lidar melhor a internacionalização, pois edições em línguas diferentes do inglês possuem uma cobertura melhor da cultura local (LEHMANN *et al.*, 2015). Outro problema, apontado pelo próprio site da DBpedia, é a duplicação de elementos ontológicos que prejudicam a extração de informações, como no caso do local de nascimento de uma

[38] Ranking do Portal Alexa, disponível em <http://www.alexa.com/topsites>. Acesso em 05/05/2016.
[39] Página de estatísticas da DBpedia, atualizada em outubro de 2015: <http://wiki.dbpedia.org/dbpedia-dataset-version-2015-10>.

pessoa que atualmente é identificada pelos elementos ontológicos *birthplace* ou *placeofbirth*. Lenman *et al.* também apontam para a possibilidade da própria DBpedia servir de base para corrigir erros da sua fonte, a Wikipedia, detectando, principalmente, inconsistência de informações entre seus artigos ou até dentro do próprio artigo, por exemplo, a data da morte de uma pessoa deve ser maior que o seu nascimento.

Figura 9 – Arquitetura da DBpedia



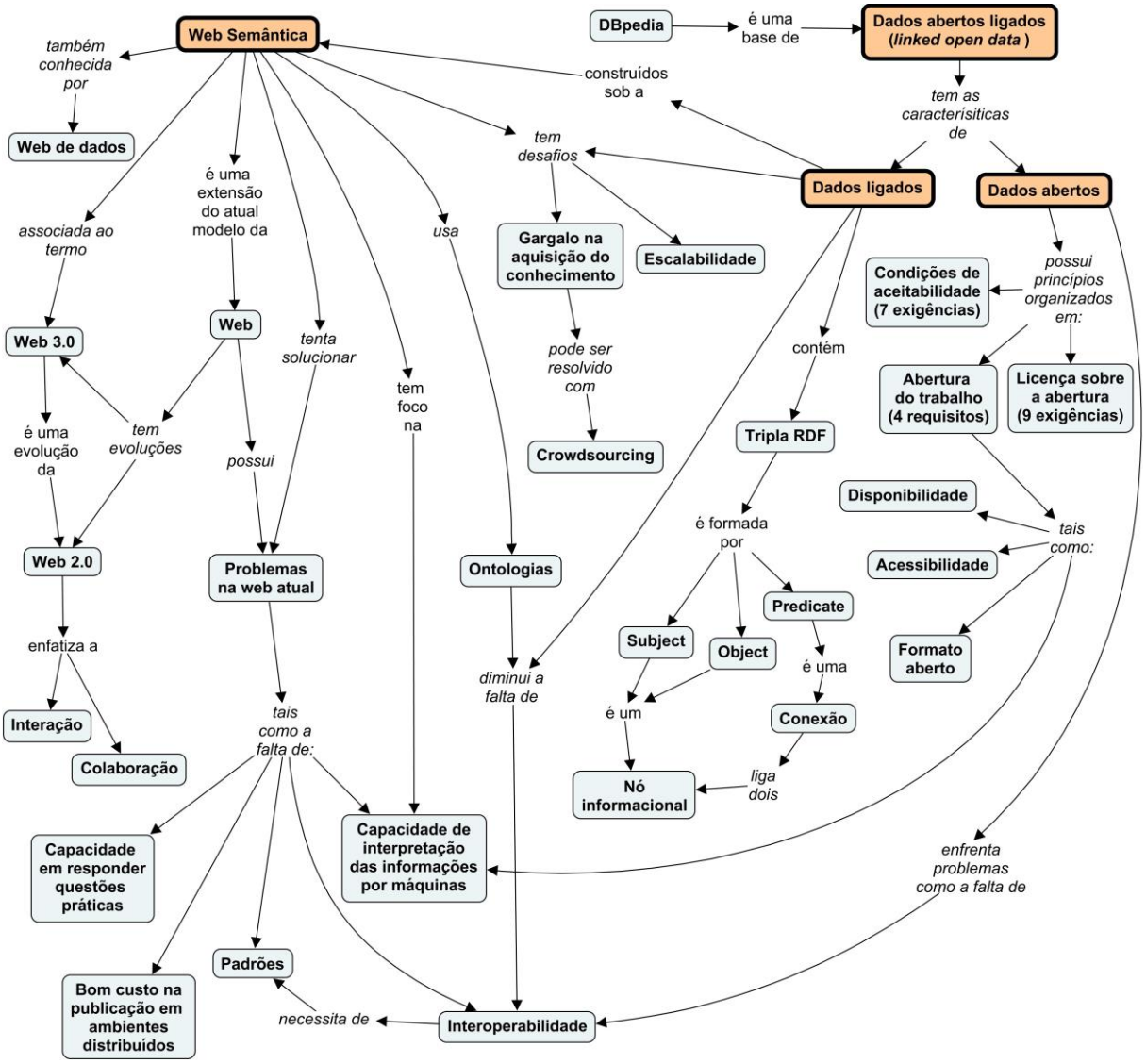
Fonte: http://wiki.dbpedia.org/about/about-dbpedia/architecture

2.3.6 Considerações finais da seção

De modo geral, a Web Semântica foi desenvolvida para resolver: a dificuldade em responder questões práticas ou específicas, a falta de padrões e a falta de capacidade para interpretação de dados via computacional. Em função disso, a web evolui no sentido de aumentar a sua capacidade semântica, ou seja, seguir as indicações das 5 estrelas de Berners-Lee. Assim, no futuro se espera que a web seja composta majoritariamente por dado abertos ligados, disponíveis em formatos não proprietários, identificados por URIs para que sejam recuperados de forma unívoca, e efetivamente ligados de acordo com a suas relações semânticas. Um dos grandes empecilhos para essa revolução é baixa interoperabilidade, que poderá ser vencida com a efetiva adoção de ontologias que se encarregam em definir a natureza das informações e suas relações, além de classificá-las e evitar ambiguidades. As

ontologias são capazes de integrar bancos de dados heterogêneos e permitir que os sistemas informáticos funcionem de forma integrada mesmo que em diferentes interfaces.

Figura 10 – Mapa conceitual com alguns relacionamentos abordados na seção 3: Web Semântica



Fonte: Elaboração própria

Apesar de problemas detectados no Portal da Transparência e na aplicação da Lei de Acesso a Informação, discutidos na seção 2.3.2, essa iniciativa é promissora se for devidamente integrada a ideia dos dados abertos ligados, com adoção do modelo das 5 estrelas de Berners-Lee e uso de ontologias adequadas. Isso permitirá um grande avanço social, considerando que muitas ferramentas de acesso à informação poderão ser desenvolvidas em função dessa nova organização de informação e conhecimento.

O mapa conceitual da Figura 10 apresenta alguns relacionamentos importantes abordados nessa seção sobre Web Semântica, destacando, em cor alaranjada e espessura maior, alguns conceitos relevantes para a presente tese. Entre as várias proposições existentes no mapa, destacam-se aquelas que caracterizam a Web Semântica revelando seus desafios e problemas atuais e comparando-a com a web. Outras proposições explicitam o termo *linked open data* separando-o em dados ligados e dados abertos. Destaque para os *crosslinks* que revelam a interoperabilidade como um problema tanto da Web Semântica quanto para os dados abertos, e propõe o uso de ontologias e dados ligados como elementos para reduzir esse problema. Além disso, a capacidade de interpretação das informações por máquina é o foco principal da Web Semântica e também é um dos quesitos para os dados abertos.

2.4 Recuperação de informação e conhecimento

Essa seção discute uma das áreas mais importantes e antigas da CI, a recuperação de informação (RI), conhecida internacionalmente por *information retrieval* e existente desde os anos 50. É interessante observar que foi a CI que se desenvolveu a partir das exigências da área da RI (WERSIG; NEVELING, 1975) e que o termo 'recuperação de informação' é possivelmente um dos mais importantes no campo da Ciência da Informação (CI) (CAPURRO; HJØRLAND, 2003). Nos estudos de Pinheiro (2007), a disciplina nomeada como Sistemas de Recuperação da Informação é destacada como a mais importante dentro da Ciência da Informação, pelo critério de número de artigos publicados.

Essa seção é organizada da seguinte forma: a primeira subseção dá uma visão geral da RI; a segunda subseção aborda a interação, o comportamento e a cognição na RI; a terceira subseção trata da recuperação de conhecimento e suas relações com a RI; a quarta subseção apresenta a visualização de informação, conhecimento e domínio no contexto da RI; a quinta subseção trata da relevância e avaliação da qualidade da informação; a sexta subseção apresenta os principais modelos de RI; finalmente, a sétima subseção discute algumas taxonomias usadas para classificar modelos de RI.

2.4.1 Visão geral da RI

Vannevar Bush, nos anos 40, já sinalizava que a capacidade humana em gerar novas informações era muito maior do que a sua capacidade em armazená-la de forma a permitir futuras consultas (BUSH, 1945). Com isso ele estava alertando sobre a necessidade de

melhorar a recuperação de informação e conhecimento. Em função disso ele chegou a imaginar um mecanismo específico para esse processo, o Memex (contração das palavras *memory* e *index*). Em 1951, Calvin Mooers, pioneiro na área da Ciência da Informação, cunhou o termo *information retrieval,* definindo-o como uma área que abrange os aspectos intelectuais da descrição das informações e sua especificação para a pesquisa, e também quaisquer sistemas, técnicas ou máquinas que são utilizadas para realizar a operação (SARACEVIC, 1999). Nos anos 60, a nova área chegou a ser entendida como núcleo da CI por vários autores (ARAÚJO, 2014).

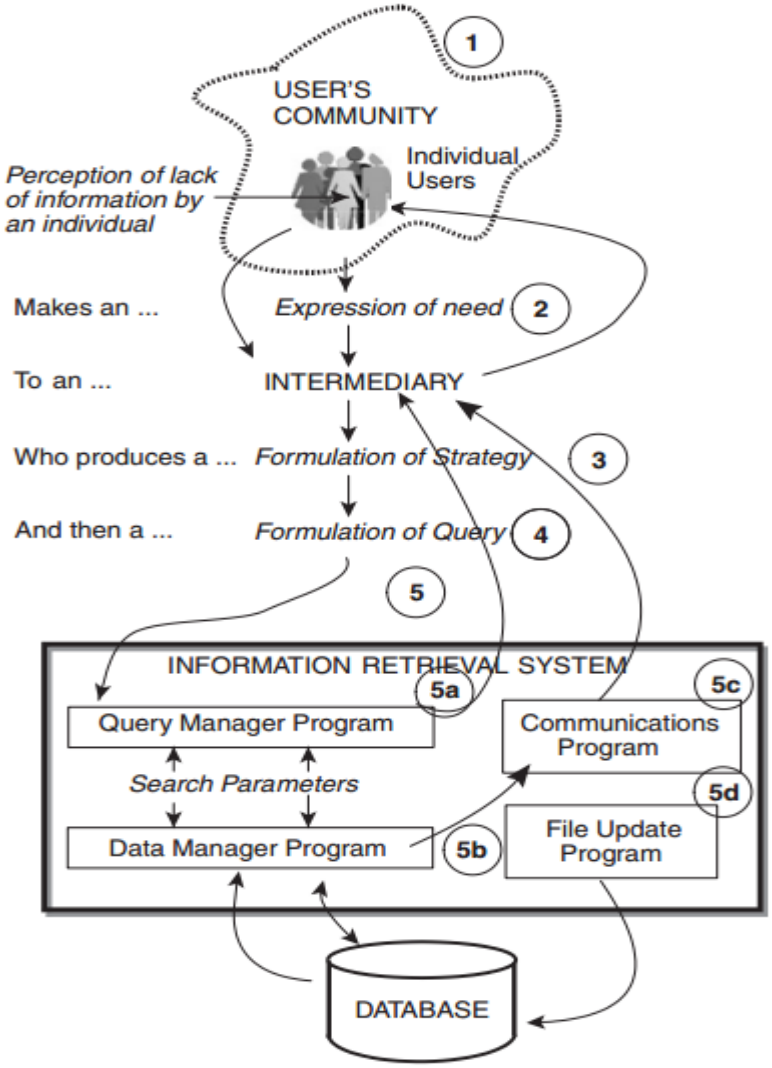
Saracevic (1999), atualizou a definição de Mooers, acrescentando a questão da interação com usuários e os aspectos contextuais: cognitivo, afetivo e situacional. Robins (2000) também retratou essa nova visão argumentando que variáveis anteriormente não consideradas nesse processo pelo modelo tradicional, como o ambiente e conhecimento do usuário, seus objetivos, intenções e crenças, começaram a ganhar força com o surgimento de novos modelos. Ainda sim, existem definições bem sucintas que não explicitam a preocupação com o usuário, tal como o glossário The Information Arquitecture (HAGEDORN, 2000), onde RI é o estudo de sistemas de indexação, busca, armazenamento e recuperação de conteúdo. Ou ainda, Salton e McGill (1983) que afirmam que a recuperação de informação tem foco na representação, armazenamento, organização e acesso de itens de informação.

Contudo, a área se desenvolveu e são predominantes as definições de RI com visão ampla envolvendo, simultaneamente, sistemas e usuários. Por exemplo, Baeza-Yates e Ribeiro-Neto (2011) defendem que o principal objetivo de um sistema de IR é recuperar todos os itens que são relevantes a uma dada consulta do usuário, e o menor número possível de itens menos relevantes. Os autores também destacam que, recentemente pesquisas em RI também incluem modelagem, busca na web, classificação de texto, arquitetura de sistemas, interface de usuários, visualização de dados, filtragem e linguagens.

A Figura 11 apresenta um modelo geral de um sistema de recuperação de informação. Observa-se nessa figura, que a recuperação de informação ocorre no âmbito de um sistema cíclico, onde usuários de uma comunidade (*user's community*), ao perceberem uma falta de informação – passo 1 – expressam uma necessidade (*expression of need*) para um agente intermediário (*intermediary*) – passo 2 – aqui também chamado de mediador. Com a popularização de RI interativas, principalmente enquanto interfaces web, o papel do mediador tem sido reduzido, mas ele é visto com importância, por ter maior experiência na formulação de estratégias de recuperação. Uma vez definida a estratégia de recuperação (*formulation os*

strategy) – passo 3 – ela é materializada na forma de uma consulta (*formulation of query*) – passo 4. A consulta é submetida ao sistema de recuperação de informação (*information retrieval system*) – passo 5, que retorna uma lista de respostas ao usuário, tipicamente uma lista de documentos ranqueados em ordem decrescente de relevância. Se o mediador atua nesse processo, ele também realiza uma avaliação dos resultados da consulta, antes de entregá-los ao usuário, tendo em vista o seu melhor entendimento das necessidades de informação, quando comparado ao que é capaz de ser inferido pelo sistema de recuperação de informação. O mediador pode refazer a consulta, sem encaminhar os resultados ao usuário, a fim de obter novos resultados que sejam mais relevantes.

Figura 11 – Visão geral de um sistema de RI



Fonte: Meadow *et al.* (2007)

Do ponto de vista técnico do processo de recuperação, parte inferior da Figura 11, destacam-se o sistema automatizado de RI propriamente dito (*information retrieval system*), e sua base de dados (*database*). O sistema automatizado de recuperação de informação é composto por:

(i) Gerenciador de consultas (*query manager program*), que transforma os termos da consulta em comandos adequados à forma de organização interna dos dados – passo 5a;

(ii) Gerenciador de dados (*data manager program*), que realiza as buscas na base de dados – passo 5b;

(iii) A interface com o mediador (*communications program*) – passo 5c; e

(iv) O atualizador/organizador da base de dados (*file update program*)– passo 5d.

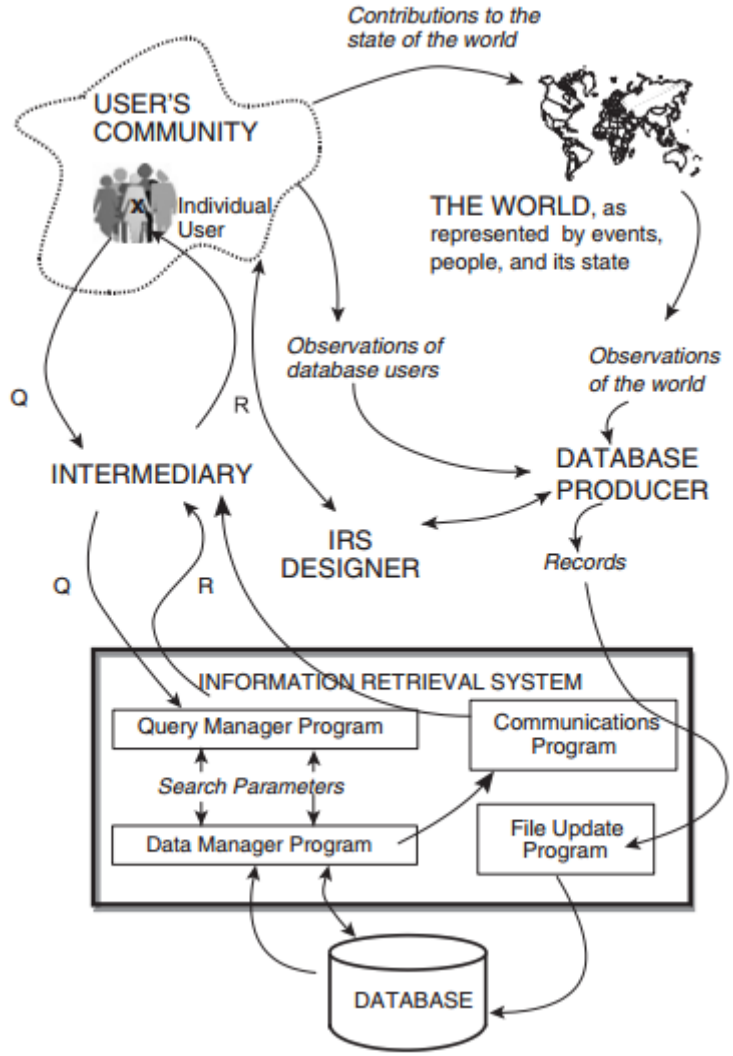
Os sistemas de RI são instrumentos essenciais para o tratamento da explosão informacional, principalmente após o advento da web. Atualmente os sistemas de RI podem oferecer velocidade de processamento que, devido às peculiaridades das bases de dados, principalmente quanto à escalabilidade e distribuição, é necessária para a absorção do grande volume de consultas. Contudo, ao longo da história, os sistemas de RI foram evoluindo e sendo desenvolvidos sob diferentes abordagens da RI. Saracevic (1999), Baeza-Yates e Ribeiro-Neto (2011) destacam duas abordagens para a RI, onde apesar de ambas terem igual interesse na recuperação, possuem diferentes perspectivas, diferenças conceituais e organizacionais:

- **RI com abordagem centrada em sistemas**: desde os anos 50 e durante várias décadas seguintes foi a única abordagem existente. Ela é baseada em algoritmos e segue um modelo que não dá relevância plena à interação com o usuário. Ou seja, ela está interessada em evoluir os componentes na parte inferior da Figura 11: o sistema de RI (*information retrieval system*) e os bancos de dados (*database).*
- **RI com abordagem centrada no usuário**: se iniciou a partir dos anos 80. Usa modelos interativos e cognitivos considerando e até dando ênfase no lado humano. Ou seja, ela está interessada em evoluir os componentes da parte superior da Figura 11.

Outro enfoque classificatório que pode ser dado a RI é conforme a sua relação com os três paradigmas da CI, físico, cognitivo e social, descritos por Capurro (2003), na seção 2.1.2.

- **RI no Paradigma Físico**: o usuário não é valorizado no processo de recuperação de informação que é mostrado como um processo mecânico onde existe, de um lado, um sistema de RI com uma base de dados, do outro, o usuário que tende a ter o seu desejo informacional não interpretado corretamente e, no centro, o profissional da informação (mediador da Figura 11) que tenta fazer a ligação entre os dois lados (ALMEIDA *et al.*, 2007). O paradigma físico está diretamente relacionado à abordagem centrada em sistemas, de Saracevic (1999) e Baeza-Yates e Ribeiro-Neto (2011), discutida nessa subseção.
- **RI no Paradigma Cognitivo**: Almeida *et al.* (2007) destacam a consideração dos modelos mentais dos usuários em alguns enfoques da RI, utilizando abordagens cognitivas e centradas no processo interpretativo dos usuários, observando as suas características fenomenológicas e individuais, "valorizando assim tentativas de inclusão das dimensões semânticas e pragmáticas nos sistemas de RI, com o intuito de possibilitar uma melhor gestão de informações a partir da análise de como as informações são compreendidas pelos usuários" (p. 22). Nesse sentido, a atividade do mediador humano, representado na Figura 11, tenta desenvolver essa abordagem em sistemas automatizados de baixa tecnologia.
- **RI no Paradigma Social**: Almeida *et al.* (2007) descrevem que em alguns enfoques da RI o foco é na "[...] recuperação dos elementos subjetivos dos usuários para a definição do desenho dos sistemas de recuperação, considerando sua visão de mundo" (p. 22). Portanto, em enfoques da RI aderentes ao paradigma social, a CI "[...] volta-se para um enfoque interpretativo, centrado no significado e no contexto social do usuário e do próprio sistema de recuperação da informação" (p. 22). Essa abordagem é melhor representada na expansão da Figura 11, conforme ilustra a Figura 12. O produtor da base de dados é explicitamente indicado como parte do sistema de recuperação de informação, bem como as interações desses com a comunidade de usuários, e as interações de ambos com o mundo, seus eventos, pessoas e estados. Também é representado nesse modelo o projetista do sistema de recuperação de informação, que precisa considerar as condições de sua comunidade de usuários e das bases de dados que poderá utilizar juntamente com o sistema de RI. Trata-se, portanto, de um modelo que considera o elevado grau de interação entre usuários e bases de dados, para a construção de sistemas de RI.

Figura 12 – Visão ampliada do ciclo de RI



Fonte: Fonte: Meadow *et al.* (2007)

2.4.2 Interação, comportamento e cognição na RI

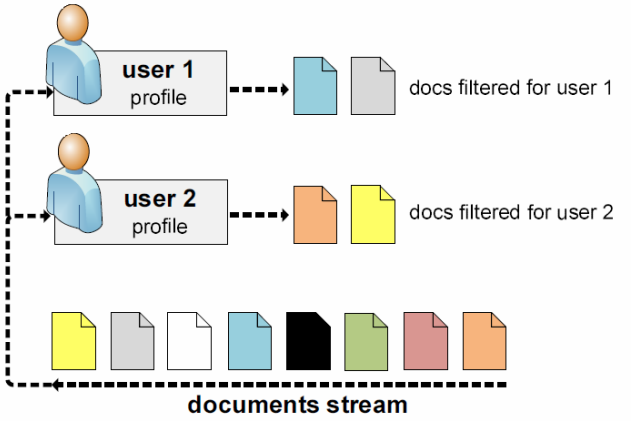
Em decorrência das preocupações da RI em direcionar o foco para o usuário, há cada vez mais tendência para estudos e desenvolvimentos de áreas relacionadas a RI que consideram fortemente a participação do usuário, tais como, a RI interativa, o comportamento informacional com as subáreas *information seeking behaviour* e *information search behaviour*, e finalmente a RI cognitiva. Essas áreas são abordadas nessa subseção.

2.4.2.1 RI interativa

Processos interativos na RI tem grande dependência do usuário e estão ficando cada vez mais frequentes. A Figura 12 representa um ciclo interativo formado entre a comunidade de usuários (*user's community*) e o sistema de RI (*information retrieval system*), mesmo com a presença do mediador (*intermediary*). Na medida em que o usuário tem contato direto com o sistema, o processo de busca e seleção de documentos recuperados pode ficar ainda mais interativo. Por exemplo, a Figura 13 mostra um processo RI interativo onde os usuários '*user 1*' e '*user 2*' filtram o resultado da recuperação de informação interagindo diretamente com o resultado da busca fornecido pelo sistema de RI, representado por um fluxo de documentos (*documents stream*).

Nesse sentido, existe uma área de estudo denominada *interactive information retrieval* (IIR) e tem como foco principal as pessoas. Ferramentas de IIR permitem explorar a informação e a descoberta de conhecimento de forma mais eficaz e eficiente (ZHANG, 2008). A IIR é a composição da RI com a área interação humano-computador, ou *human computer interaction* (KELLY; SUGIMOTO, 2013). Dessa forma, Belkin (2008) e Kelly e Sugimoto (2013) alertam para o desafio de abordar a natureza inerentemente interativa da RI, ou seja, é necessário considerar as interações do usuário como um processo central da RI. Por outro lado, os autores reconhecem que isso exigirá esforços conjuntos de pesquisadores que terão que se abdicar dos modelos estritamente formais de RI, em favor de modelos realistas e úteis.

Figura 13 – Recuperação via filtragem pelo usuário



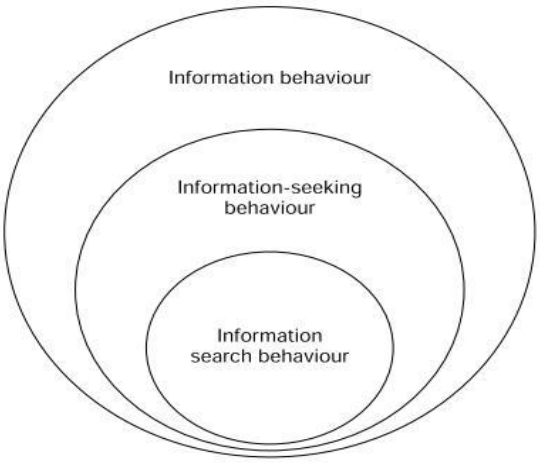
Fonte: Baeza-Yates e Ribeiro-Neto (2011)

2.4.2.2 Comportamento informacional

A RI também está relacionada a uma área denominada de *information seeking behavior*, também conhecida por *information seeking and retrieval* (IS&R) (INGWERSEN; JÄRVELIN, 2005). Porém, essa área é ampla e altamente dependente do contexto como os estados cognitivos e afetivos do usuário, ambientes sociais, culturais e organizacionais entre outros (SARACEVIC, 2010). Saracevic também destaca outra área similar: *information search behavior*, que é um subconjunto da *information seeking behaviour* e, no contexto da CI, refere-se aos processos utilizados para capturar e recuperar informações em diferentes sistemas de informação, sendo a parte mais empírica e pragmática da *information seeking behaviour*, onde muitos estudos estão voltados para a melhoria da interface, motores de busca e das interações humano-computador. Segundo Wilson (2000), *information search behaviour* acontece num nível micro e engloba todas as interações do usuário com os sistemas de informação, desde a utilização do mouse até um nível mais intelectual com a utilização de uma estratégia de busca booleana ou mesmo a determinação de critérios para julgar qual de dois livros selecionados é mais útil.

Ambas as áreas de pesquisa, *information seeking behaviour* e *information search behaviour* pertencem a uma área maior denominada de comportamento informacional, ou *information behavior* ou *human information behavior*, que trata de forma mais ampla o comportamento do usuário, como mostra a Figura 14, referindo-se a um conjunto de processos e estratégias executadas dinamicamente por pessoas para buscar informação sobre suas questões (SARACEVIC, 2010).

Figura 14 - Composição dos comportamentos de busca



Fonte: Wilson (1999)

Para Ingwersen e Järvelin (2005), comportamento informacional também inclui a área RI interativa, a as áreas já citadas *information seeking behaviour* e *information search behaviour*, além de lidar com a geração, comunicação e utilização da informação. Comportamento informacional é um conceito de interesse de campos como a Psicologia, a Sociologia e Ciência Política, e é altamente dependente do contexto, envolvendo várias motivações para a busca da informação, vários estados cognitivos e afetivos, vários ambientes sociais, culturais e organizacionais, várias características, valores e modos de vida, e assim por diante (SARACEVIC, 2010). Segundo Wilson (2000) é o comportamento humano em relação às fontes e canais de informação, incluindo busca de informação ativa e passiva, e uso da informação. Segundo Fisher e Julien (2009) o estudo do comportamento informacional foca nas necessidades de informação das pessoas, como elas procuram, gerenciam, entregam, e utilizam informações, tanto propositadamente quanto passivamente, nos papéis variados que compõem suas vidas cotidianas. Araujo (2014), destaca a realização de "[...] estudos sobre 'práticas informacionais', voltados para o estudo da ligação entre aspectos informacionais socioculturais (formas coletivas de se relacionar com a informação, critérios coletivos de relevância, necessidade etc.) e os comportamentos informacionais individuais" (p. 13).

Case (2006) afirma que a área comportamento informacional não deverá mais sofrer grandes modificações no campo teórico por já possuir indicações de maturidade. É possível para o pesquisador compreender essa área como um processo inerente ao ser humano aprendiz que na medida em que tem necessidade de informação executam ações de busca, uso e transferência de informação (GASQUE; COSTA, 2010). Finalmente, Saracevic (2010) afirma que comportamento informacional é forte em modelos e teorias e tem seguido uma linha de avançar na direção de outras disciplinas, particularmente, a Psicologia, Comunicação e Filosofia.

2.4.2.3 RI cognitiva

A RI cognitiva, *cognitive information retrieval* (CIR), é relacionada aos processos de interação homem-computador que estão embutidos no contexto social e vida cotidiana do indivíduo (SPINK; COLE, 2005). Spink e Cole ainda sinalizam que CIR é uma área interdisciplinar de pesquisa que inclui CI, Ciência da Computação, Ciência Cognitiva, Interação Homem-Computador, fatores humanos e outras disciplinas relacionadas à informação. Os autores também argumentam que essa é uma área importante a ser considerada no desenvolvimento de sistemas de RI.

Spink e Cole abrem reflexão para analisar a distinção entre abordagem de sistema e abordagem cognitiva fazendo uma comparação com a diferença entre busca de informação e uso de informação, e ainda entre dado e informação. A partir dessa reflexão os autores indicam que a contextualização da recuperação de dados em busca e uso de informação na interação entre usuário e sistema é talvez o caminho para transformar o sistema de RI, de um sistema de aquisição de dados em um sistema de aquisição de informações para o usuário.

Sobre essas duas abordagens, Ingwersen e Järvelin (2005) destacam que a RI cognitiva enfrenta problemas metodológicos precisando ser convincente sobre a generalização dos resultados, em particular aqueles derivados de estudos orientados para o usuário. Isso porque há uma diferença fundamental: a RI cognitiva precisa controlar a questão humana que é dependente de inúmeras variáveis. Por outro lado, os autores destacam que a RI pode ser vista como uma ciência ou tecnologia sobre como aumentar o desempenho no acesso às informações em documentos em vários meios de comunicação, e a RI cognitiva aparece como uma disciplina fundamental nesse processo.

2.4.3 Recuperação de conhecimento

Para investigar, definir e situar melhor a recuperação de conhecimento, um caminho é compará-la com outros tipos de recuperação mais conhecidos. Dessa forma, Yao *et al.* (2007) classificam os sistemas de recuperação em geral em três categorias: (i) sistemas de recuperação de dados (*data retrieval systems* – DRS), adequados para armazenamento de recuperação de dados estruturados, tal como os sistemas gerenciadores de bando de dados; (ii) sistemas de recuperação de informação (*information retrieval systems* – IRS), busca as informações solicitadas pelo usuário, tal como as máquinas de busca da web, porém, o usuário é quem deve ler e analisar quais são os documentos relevantes que contém conhecimento útil; (iii) sistemas de recuperação de conhecimento (*knowledge retrieval systems* – KRS), que tem o foco no conhecimento que fornecem, porém, ainda são um grande desafio. O Quadro 2 compara e sintetiza esses três tipos de recuperação

Por intermédio de critérios comparativos (casamento de busca, inferência, modelo, consulta, organização, representação, armazenamento e resultado recuperados), Yao *et al.* (2007) distinguiram os três tipos de recuperação, apresentados no Quadro 2. Porém, é possível que um sistema tenha elementos mesclados de uma ou outra coluna, não sendo puramente de um só tipo.

Existem controvérsias quanto ao escopo de atuação da recuperação de conhecimento. Por exemplo, Yao *et al.* (2007) observaram que alguns autores consideram que a recuperação de conhecimento é um processo da RI. Há quem considere que a recuperação de conhecimento é uma forma de inferência e acontece sobre uma base de conhecimento. Campos, Souza e Campos (2003) reconhecem a recuperação de conhecimento quando realizada por sistemas baseados em conhecimento, pois "[...] ao contrário do convencional processamento de dados, têm elementos que são mais do que dados isolados; são conceitos que descrevem objetos e suas propriedades" (p. 11). Como técnicas para avançar nos métodos para a recuperação de conhecimento, Martin e Eklund (2000) propuseram um ambiente baseado no uso de linguagens de metadados, tais como a RDF e linguagens de ontologias. Já nessa época, ano 2000, eles reconheceram o problema da escalabilidade da web e também a falta de semântica nas informações disponíveis na web.

A recuperação de conhecimento também pode ser vista como um processo que acontece com um indivíduo na sua esfera privada "[...] com o objetivo de lembrança de um dado conceito, previamente assimilado e associado a sua estrutura de conhecimentos, por meio de métodos que promovam a recuperação eficiente e efetiva destes conceitos, dentro de um domínio de conhecimento" (PONTES JUNIOR; CARVALHO; AZEVEDO, 2013, p. 14). Os autores ainda sinalizam que entender como o conhecimento é organizado permite o estabelecimento de "[...] correlações entre os processos de recuperação de conhecimento com os de recuperação de informação, objetivando a melhoria do desempenho desses últimos" (p. 14). Nessa ótica, a recuperação de conhecimento é dependente do usuário, de seus conhecimentos anteriores e como as informações recuperadas irão se relacionar com o seu conhecimento prévio.

Quadro 2 – Comparação entre recuperação de dados, de informação e de conhecimento

	Recuperação de dados	Recuperação de informação	Recuperação de conhecimento
Casamento da busca	booleano	parcial ou melhor	parcial ou melhor
Inferência	dedutiva	indutiva	dedutiva, indutiva, raciocínio associativo, raciocínio analógico
Modelo	determinístico	estatístico e probabilístico	semântico, de inferência
Consulta	linguagem artificial	linguagem natural	estrutura de conhecimento, linguagem natural
Organização	tabela, índice	tabela, índice	unidade de conhecimento, estrutura de conhecimento
Representação	número, regra	linguagem natural, linguagem de marcação	grafo de conceitos, lógica de predicados, regra de produção, estrutura, rede semântica, ontologia
Armazenamento	banco de dados	coleções de documentos	base de conhecimento
Resultados recuperados	conjunto de dados	seções ou documentos	um conjunto de unidade de conhecimento

Fonte: adaptado de Yao *et al.* (2007) extensão de Van Rijsberngen (1979)

2.4.4 Visualização na RI

Uma das etapas de um processo de RI é o relacionamento com a informação recuperada por parte do usuário que demandou a consulta. Esse relacionamento pode ocorrer de forma simples com o auxílio do formato textual, ou de uma forma mais elaborada, com auxílio de imagens, mapas esquemas etc., para proporcionar ao usuário a obtenção de *insights* sobre a informação recuperada. Essa subseção apresenta algumas áreas de estudo que têm investigado a melhoria desse processo de apresentação da informação ao usuário: visualização de informação (*information visualization*), visualização de conhecimento (*knowledge visualization*), visualização de informação na RI (*information retrieval visualization*) e visualização de domínio de conhecimento (*knowledge domain visualization*).

2.4.4.1 Visualização de informação e de conhecimento

O termo visualização refere-se à representação gráfica de dados ou conceitos, diferente do conceito antigo, que o associava a imagens mentais que as pessoas formavam enquanto pensavam (WARE, 2010). A visualização é também considerada um processo de transformação de dados, informação e conhecimento em representações gráficas para apoiar tarefas, tais como análise de dados, exploração de informação, explicação da informação, previsão de tendências, detecção de padrões, descobertas etc. (ZHANG, 2008). As visualizações estão cada vez mais presentes nas ciências e os recentes avanços na exploração do cérebro têm permitido aos pesquisadores ver quais áreas do cérebro são mais ativas quando as pessoas desempenham tarefas visuais (WARE, 2010). Normalmente as pessoas ficam mais sintonizadas com imagens e informação visual do que informação puramente textual e, a representação visual tem condições de comunicar alguns tipos de informação muito mais rapidamente e com efetividade do que outro métodos (BAEZA-YATES; RIBEIRO-NETO, 2011). O estudo de temas baseados em texto pode ser significativamente reforçado pelos benefícios da visualização da informação e espacialização (representação espacial de dados não-espaciais) (HOOK; BÖRNER, 2005). A informação é mais fácil de lembrar se ela puder ser armazenada e localizada espacialmente (MILLER, 1968; WINN, 1994 apud HOOK; BÖRNER, 2005, p. 189).

Segundo Keller e Tergan (2005), visualização de informação (*information visualization*) e visualização de conhecimento (*knowledge visualization*) são áreas de estudo distintas com desenvolvimentos próprios, apesar de ambas as áreas terem como objetivo comum a visualização de estruturas. Onde estruturas referem-se tanto elementos de conhecimento quanto de informação. A visualização de informação emergiu como um campo de estudo nos anos 90 e tem como objetivos, revelar padrões invisíveis a partir de dados abstratos e trazer novas percepções para as pessoas, e não apenas imagens bonitas (CHEN, 2013). O autor ainda enfatiza que o maior desafio é capturar algo abstrato e invisível para algo concreto, tangível e visualmente significativo.

Quadro 3 – Comparação entre visualização da informação e visualização do conhecimento

Aspectos	Visualização da informação (*information visualization*)	Visualização do conhecimento (*knowledge visualization*)
Objetivo	Uso de aplicações auxiliadas por computador em tarefas exploratórias de grande quantidade de dados para obter novos *insights*.	Uso de representações visuais para melhorar a transferência de conhecimentos entre as pessoas e melhorar a criação de conhecimento em grupos.
Benefício	Melhora o acesso à informação, recuperação e exploração de grandes bancos de dados.	Multiplica os processos de transferência de conhecimento e de comunicação entre os indivíduos usando uma ou mais representações visuais.
Conteúdo	Concentra-se em dados explícitos, tais como fatos e números.	Cuida de conhecimentos, tais como experiências, insights, instruções, suposições e aqueles que respondem a questões ‘porque’, ‘quem’ e ‘como’.
Destinatários	Auxilia um indivíduo a obter novos *insights*.	Auxilia um indivíduo ou um grupo na transferência ou criação de novos conhecimentos em ambientes colaborativos.
Influencia	Fornece novas pistas para os campos da CI, mineração de dados, análise de dados e para problemas como a exploração de informação, recuperação de informação, interação humano-computador e design de interface.	Fornece novas pistas para os campos de ciência da comunicação visual, gestão do conhecimento, e para problemas como a exploração, transferência, criação e aplicação de conhecimento; a aprendizagem, a qualidade da informação, sobrecarga de informação, design, design de interface e comunicação visual.
Proponentes	Pesquisadores normalmente possuem base na ciência da	Pesquisadores normalmente possuem base na gestão do

	computação.	conhecimento, psicologia, design ou arquitetura.
Contribuição	Orientada à inovações; pesquisadores criam métodos mais técnicos.	Orientada à soluções; também cria métodos de visualização tradicionais para resolver problemas, porém, somente na inexistência dos métodos. Proporciona integração entre áreas isoladas oferecendo estruturas teóricas necessárias para toda a área de visualização.
Raízes	Campo novo que se tornou possível com a introdução de computadores.	Apesar do termo recente, visualização conhecimento, é baseada em realizações culturais e intelectuais, por exemplo, de arquitetos e filósofos.
Significados	Usa métodos auxiliados por computador.	Usa métodos de visualização auxiliados por computador ou não.
Visualizações complementares	Combina métodos diferentes de visualização usando o mesmo meio em uma interface, acoplando-os firmemente; este conceito é chamado de múltiplas visualizações coordenadas.	Combina diferentes métodos de visualização usando um ou mais meios diferentes (por exemplo, um software, um cartaz ou um objeto físico) para ilustrar conhecimentos em perspectivas diferentes.

Fonte: adaptado de Burkhard (2005)

A visualização de conhecimento examina o uso de representações visuais para melhorar a transferência e a criação de conhecimento entre pelo menos duas pessoas (BURKHARD, 2005). Segundo Burkhard, essa definição é aceita por especialistas das áreas de visualização de informação (*information visualization*) e visualização de domínio do conhecimento (*knowledge domain visualizations)*. Keller e Tergan (2005) focam o termo visualização de conhecimento nas visualizações de estruturas que representam conhecimento conceitual. Meyer (2010) define a visualização de conhecimento como um campo de investigação com foco sobre a criação e transferência de conhecimento por intermédio de

visualizações com ou sem a ajuda de computadores, sendo ainda um mediador capaz de criar um efeito sinérgico entre várias disciplinas diferentes (MEYER, 2010).

Burkhard (2005) analisou diferenças entre as áreas visualização de informação e visualização de conhecimento. O autor observou que em ambos os casos, pesquisadores exploram habilidades inatas para processar de forma efetiva as representações visuais, mas o modo de usar essas habilidades diferem nos dois casos. Burkhard organizou as diferenças em dez aspectos, aqui sintetizadas e tabuladas, como mostra o Quadro 3.

No Quadro 3 observa-se o quanto a visualização de conhecimento é mais efetiva, do que a visualização de informação, para processos de comunicação e construção do conhecimento em meios cooperativos. Eppler e Burkhard (2004) também destacam vantagens da visualização de conhecimento, sobre a visualização de informação, benefícios cognitivos, sociais e emocionais. Os autores resumiram os pontos fortes no acrônimo 'CARMEN':

- 'C': *coordination* (coordenação) – coordenação na comunicação entre profissionais do conhecimento – benefício social;
- 'A': *attention* (atenção) – sensibilização e foco para a criação e transferência do conhecimento – benefício cognitivo;
- 'R': *recall* (recuperação) – melhoria da capacidade de memorização e, assim, fomentar a aplicação de novos conhecimentos – benefício cognitivo;
- 'M' *motivation* (motivação) – motivação dos usuários para engajá-los na interpretação e exploração de figura/gráfico – benefício emocional;
- 'E': *elaboration* – (elaboração) – maior compreensão e apreciação de conceitos e ideias e suas relações; e
- 'N': *new insights* – (novas percepções) – revelação de conexões escondidas – benefício cognitivo.

Eppler e Burkhard também citam exemplos de visualização do conhecimento: (i) esboço heurístico: criação de novos *insights* em grupo; (ii) diagramas conceituais: estruturação de informação e ilustração de relacionamentos; (iii) metáforas visuais: descrição de domínios para melhoria do entendimento; (iv) animações de conhecimento: visualizações dinâmicas e interativas; (v) mapas de conhecimento: navegação e estruturação especialista; (vi) gráficos científicos: visualização de domínio de conhecimento e estruturas intelectuais.

Dentre esses exemplos citados, é destacado aqui, devido ao contexto do presente trabalho, a estruturação de informação e ilustração de relacionamentos por intermédio de diagramas conceituais. Eppler e Burkhard explicam que eles são descrições esquemáticas de ideias abstratas com o auxílio de formas padronizadas (caixas, círculos, pirâmides, setas etc.) usadas para estruturar a informação e ilustrar relacionamentos. Além disso, segundo Huff (1990 apud EPPLER; BURKHARD, 2004), para a transferência e criação de conhecimento, os diagramas conceituais ajudam a tornar mais acessíveis os conceitos abstratos, reduzindo a complexidade das questões-chave, possibilitando a amplificação da cognição e a discussão de relacionamentos.

2.4.4.2 Visualização de informação na RI

Alguns autores contextualizam a visualização da informação na RI, referindo-se como *information retrieval visualization*. Zhang (2008) define essa área como um processo que transforma os dados abstratos invisíveis e suas relações semânticas em coleções de dados visíveis. Segundo o autor, esse processo é composto por dois componentes: (i) apresentação de informações visuais; e (ii) recuperação de informação visual. A apresentação de informações visuais é quem fornece a plataforma onde a recuperação de informação visual é executada ou realizada. Zhang ainda enumera sete benefícios do uso da visualização na RI que vão desde a utilização da capacidade perceptiva humana, passando pela redução da carga de trabalho cognitivo, até a melhoria na eficácia da recuperação. De forma resumida, os benefícios são:

(i) Oferece um ambiente ideal e natural para permitir a navegação;

(ii) Realiza a espacialização de um contexto informacional, isto é, torna um espaço de informação invisível e abstrato para um espaço visual;

(iii) Transforma a informação agregada num nível macro para uma coleção de dados disponível e acessível para os usuários;

(iv) Oferece caminhos diferenciados para o desenvolvimento de novos modelos de RI, diferentes dos modelos tradicionais;

(v) Fornece condições bastante favoráveis para a análise da informação, como por exemplo, a análise de conexões entre informações;

(vi) Abre amplas possibilidades para o desenvolvimento de abordagens para as apresentações visuais, principalmente pela espacialidade; e

(vii) Enriquece e eleva o nível da RI tornando o processo intuitivo e simples, e ainda deixando os usuários capazes de construir e realizar descobertas de conhecimento.

2.4.4.3 Visualização de domínio de conhecimento

A visualização de domínio do conhecimento, ou *knowledge domain visualizations* (KDVs) (HOOK; BÖRNER, 2005), ou ainda *knowledge domain visualization* (KDViz) (CHEN, 2013) tem proximidade com a área visualização de conhecimento, porém, específica para o mapeamento e delimitação de um domínio de conhecimento. Segundo Hook e Börner (2005), visualização de domínio de conhecimento são renderizações gráficas de dados bibliométricos concebidos para proporcionar uma visão global de um domínio particular, os detalhes estruturais de um domínio, e as características mais importantes de um domínio. Renderizações gráficas de dados bibliométricos são composições de informações publicadas ou conhecidas sobre o domínio de um conhecimento, apresentadas num formato gráfico condensado. Para os autores, visualização de domínio de conhecimento é também referida como mapas de domínio e o processo de sua criação como mapeamento de domínio.

No contexto da descoberta de conhecimento, ou *knowledge discovery*, a visualização de domínio de conhecimento pode diminuir o espaço de busca e aumentar a chance de encontrar uma linha de pesquisa promissora na investigação científica (CHEN, 2013). Chen chama atenção para o fato de que a visualização de domínio do conhecimento pode ser prejudicada em função da inexistência clara das fronteiras desse conhecimento com os demais, bem como existência de conhecimento latente, isto é, conhecimento que precisa ser descoberto ou revelado. Para essa área de estudo, ele chamou de domínio de conhecimento latente, ou *latente domain knowledge*.

Há provas convincentes da utilidade da visualização de domínio de conhecimento decorrentes dos campos da Psicologia Educacional, Ciência Cognitiva, Cartografia e Ciência da Informação (HOOK; BÖRNER, 2005). Para os autores, a visualização de domínio de conhecimento bem concebida tem a capacidade de facilitar a compreensão, lembrança, e para transmitir ao usuário a organização esquemática, geo-espacial, temporal, semântico, ou social do domínio subjacente. Hook e Börner ainda acreditam que visualizações domínio do conhecimento educacional podem ajudar com o acesso e a navegação, compreensão, gestão e comunicação dos espaços de informação em larga escala. Eles destacam que visualização de domínio do conhecimento, quando utilizada como uma interface para recuperação de informação, tem o potencial de transmitir a organização estrutural do domínio para o usuário.

Os autores ainda acrescentam que, esse conhecimento estrutural do domínio fornece o fundamento cognitivo com a qual o usuário pode associar detalhes adicionais sobre o domínio.

2.4.5 Relevância e avaliação da qualidade da informação recuperada

A relevância, determinada pelo usuário, e amplamente usada nas métricas para cálculo da qualidade da informação recuperada. Especificamente no contexto da recuperação de texto, existe um método para calcular a qualidade comparando a informação recuperada com conjuntos de dados com relevância estabelecida a priori. Essa subseção apresenta três elementos importantes desse processo: a relevância; a avaliação da qualidade da informação recuperada e a avaliação de recuperação de texto em larga escala.

2.4.5.1 Relevância

A relevância da informação recuperada é dependente do usuário e não do sistema (HJØRLAND, 2010) e é um conceito chave na CI e em particular da área da RI (SARACEVIC, 1999, 2007). Porém, a determinação da relevância de cada um dos documentos recuperados é que irá permitir a construção de um *ranking* para apresentar ao usuário o grupo de documentos numa ordem mais adequada (BAEZA-YATES; RIBEIRO-NETO, 2011). Os autores ainda afirmam que os algoritmos de ranqueamento constituem-se o núcleo dos sistemas de RI. Saracevic (1999, 2007) afirma que no contexto da CI, relevância indica uma relação e é o atributo ou critério que reflete a eficácia da troca de informações entre os usuários e sistemas de RI, com base em avaliações feitas por humanos. Por outro lado, Hjørland (2010) observa que usuários de sistemas de informação não são automaticamente competentes pra julgar relevância. Pelo seu caráter abstrato, a relevância da informação recuperada dificulta a criação de estruturas artificiais capazes de garantir que os resultados de uma busca sejam relevantes ao seu usuário (SILVA; SANTOS; FERNEDA, 2013).

Saracevic (2007) analisou 16 estudos que investigaram a variedade de fatores que influenciam na forma como os seres humanos determinam a relevância da informação, observando que por um lado ela está intimamente relacionada aos estudos do comportamento informacional e por outro ela é completamente ligada à tecnologia da informação. Saracevic

conseguiu agrupar os vários fatores que determinam a relevância em seis tipos distintos, quando se trata de interação entre pessoas, informações e tecnologia:

- **Conteúdo**: tema, qualidade, profundidade, escopo, atualidade, atendimento, clareza;
- **Objeto**: características dos objetos de informação, por exemplo, tipo, organização, representação, formato, disponibilidade, acessibilidade, custos;
- **Validade**: precisão das informações fornecidas, reputação, confiabilidade das fontes, comprovação;
- **Utilização**: adequação com a situação ou tarefas, usabilidade, urgência, valor no uso;
- **Correspondência cognitiva**: compreensão, novidade, esforço mental;
- **Correspondência afetiva**: respostas emocionais à informação, diversão, frustração, incerteza;
- **Correspondência com a crença**: crédito pessoal atribuído à informação, confiança.

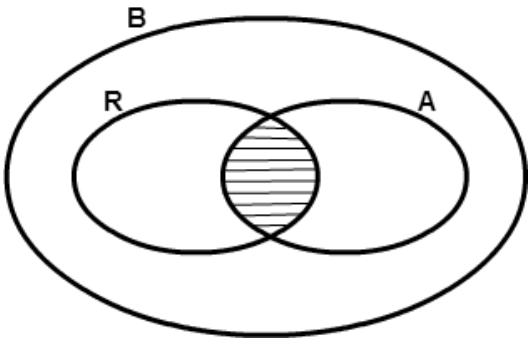
Saracevic alerta que esses critérios não são independentes, ou seja, usuários podem aplicar vários deles simultaneamente enquanto decidem sobre a relevância de uma determinada informação. O autor ainda sistematizou fatores que afetam um julgamento sobre a relevância a partir de outras pesquisas, além de desenvolver, em seu artigo, um amplo debate sobre esses fatores e suas derivações. Saracevic concluiu que, apesar das grandes mudanças nos sistemas de recuperação de informação e em ritmo acelerado, a relevância é atemporal e sempre haverá um grande foco sobre seu estudo.

2.4.5.2 Avaliação da qualidade

Cleverdon (1962), em seu projeto Aslib-Cranfield, desenvolvido na Universidade de Cranfield, investigou a eficiência de sistemas de indexação, e é considerado o precursor das medidas de relevância *precision* (precisão) e *recall* (revocação). Essas medidas foram e ainda são amplamente usadas na avaliação da qualidade dos sistemas de RI. A força delas decorre do fato de envolvem usuários como os juízes da eficácia do desempenho e, por outro lado, o ponto fraco é o mesmo, ou seja, pelo fato de depender do julgamento de usuários, possui todos os perigos da sua subjetividade e variabilidade (SARACEVIC, 1999).

Quanto ao papel dos índices de precisão em um processo de busca e recuperação de informação, Araújo Junior (2007) destaca que eles devem "[...] dar noção exata se o que está sendo recuperado na base de dados é útil ao usuário" (p. 83). O autor também observa que, apesar de trazer imprecisões, a sua utilização deve servir de ponto de partida para melhoria contínua dos sistemas de RI, mas, tudo começa pela compreensão clara da demanda informacional. Mesmo após quatro décadas, Araújo Junior (2007), Kelly (2009), Baeza-Yates e Ribeiro-Neto (2011) e Kelly e Sugimoto (2013) destacam as medidas 'precisão' e 'revocação' como bem estabelecidas e aceitas pela comunidade para a RI clássica, ou seja, aquela que não envolve interação do usuário.

Figura 15 – Medidas: precisão e revocação



Fonte: Elaboração própria

A Figura 15 representa as medidas de precisão e revocação, juntamente com as equações (5) e (6), onde:

B: é a base de documentos;

R: é o conjunto formado por todos os documentos relevantes para o usuário;

|R|: é a quantidade dos documentos relevantes, presentes na base;

A: conjunto formado por todos os documentos recuperados na busca;

|A|: é a quantidade de documentos recuperados;

R ∩ A: é a intersecção entre os conjuntos dos documentos relevantes pra o usuário e dos documentos recuperados na busca;

|R ∩ A|: é a quantidade de documentos presente no conjunto dos relevantes e recuperados.

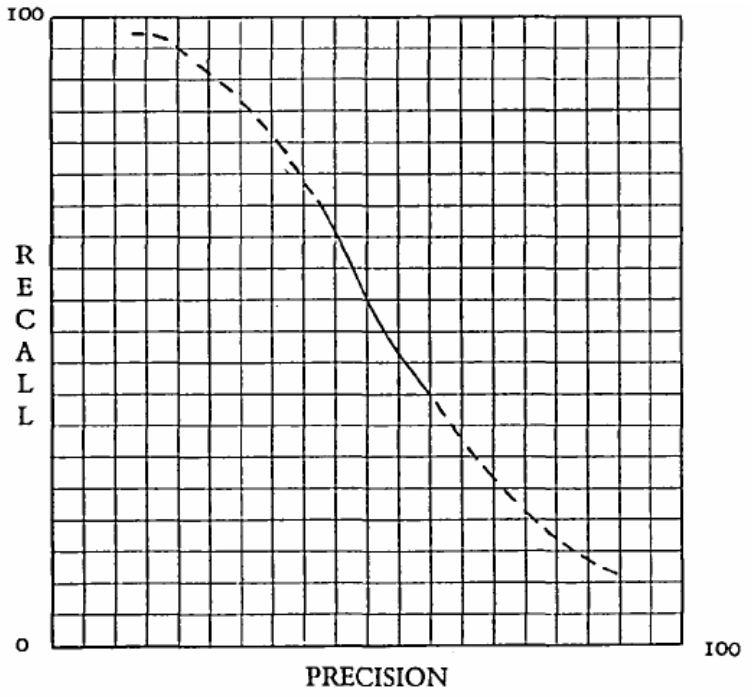
$$\textbf{revocação} = \frac{|R \cap A|}{|R|} \qquad (5)$$

$$\text{precisão} = \frac{|R \cap A|}{|A|} \quad (6)$$

Conforme a equação 5, a revocação de uma operação de recuperação de informação é mensurada pela razão entre a quantidade de documentos relevantes ao usuário e que também foram recuperados pela busca, dividida pela quantidade total de documentos relevantes na base. De acordo com a equação 6, a precisão de uma operação de recuperação de informação é mensurada pela razão entre a quantidade de documentos relevantes ao usuário e que também foram recuperados pela busca, dividida pela quantidade total de documentos recuperados pela busca.

Kelly e Sugimoto (2013) caracterizam ‘precisão’ e ‘revocação’ como medidas de desempenho, ou seja, equivalem ao quão bem sucedido um usuário foi em realizar uma tarefa de recuperação de informação junto a um sistema de TI, e ainda são derivadas da medida de relevância. A que a precisão é inversamente proporcional a revocação, ou seja, quanto maior for uma menor será a outra. Cleverdon (1972), de forma empírica, foi o primeiro a observar essa relação entre ‘precisão’ e ‘revocação’, conforme mostra o Gráfico 2.

Gráfico 2 – Relação típica entre as medidas ‘precisão’ e ‘revocação’

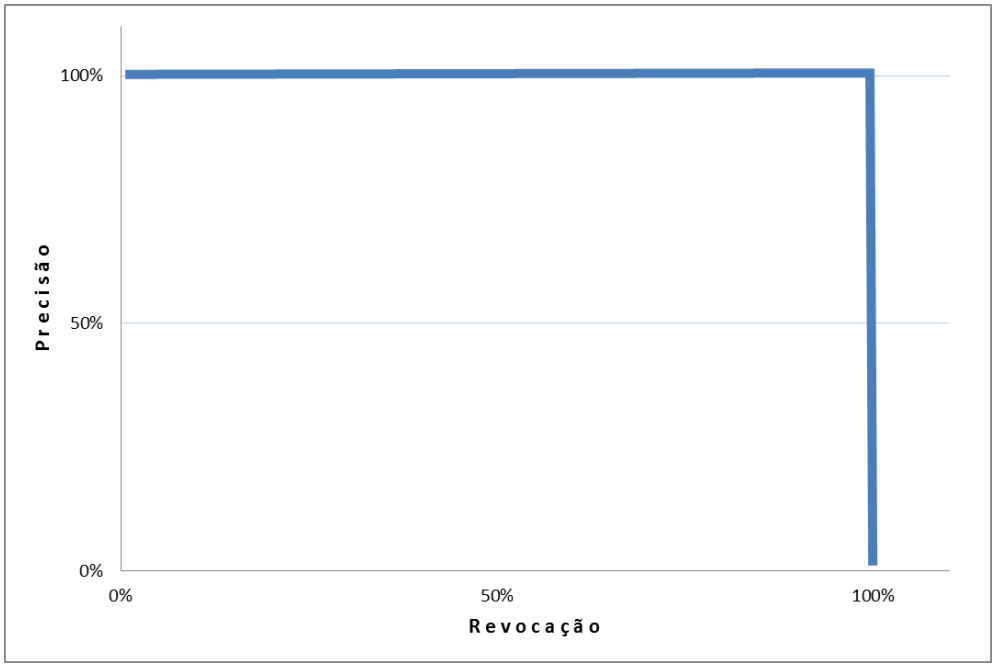


Fonte: Cleverdon (1972)

Gordon e Kochen (1989), também de forma empírica, apresentaram um gráfico próximo do resultado obtido por Cleverdon. Vários outros autores observam essa relação inversa entre ‘precisão’ e ‘revocação’, sendo que Araújo Junior (2007) destaca que em função disso, há necessidade de gestão da precisão que é apoiada em dois elementos básicos: (i) controle da revocação e da exaustividade[40], com o objetivo de aumentar o índice de precisão no processo de busca e recuperação de informação; e (ii) aperfeiçoamento contínuo da interação dos usuários com os sistemas, para aperfeiçoar a análise das medidas de precisão.

Buckland e Gey (1994) também questionam esse típico resultado entre a relação de ‘precisão’ e ‘revocação’, sugerindo, em seu experimento, a adoção de outra estratégia para minimizar essa relação inversa. Eles mostraram melhoria simultânea em ambas as medidas por intermédio de dois estágios na estratégia de recuperação: uma busca inicial, dando foco maior em uma alta revocação, em seguida, outra busca no conjunto de informações recuperadas para melhorar a precisão. Eles também propõem continuidade desse processo em vários estágios de recuperação. Essa recuperação em múltiplos estágios foi também sugerida por Porter (1983 apud BUCKLAND; GEY, 1994) e Buckland *et al.* (1992).

Gráfico 3 – Situação hipotética do relacionamento entre ‘precisão’ e ‘revocação’ em uma recuperação perfeita



Fonte: Adaptado de Buckland e Gey (1994)

[40] Exaustividade diz respeito a capacidade do sistema em indexar um documento em profundidade sendo uma variável dependente da política de indexação adotada.

A melhor situação possível para as duas medidas seria o valor absoluto de índice igual a 100% para cada uma medida de forma independente. Buckland e Gey (1994) apresentam como seria uma situação hipotética como essa, no Gráfico 3.

Como uma única medida para representar ‘precisão’ e ‘revocação’, calcula-se a média harmônica entre elas, conforme ilustra a equação (7), chamada de F-measure no contexto da RI (BAEZA-YATES; RIBEIRO-NETO, 2011):

$$\text{F-measure} = 2 \times \frac{(\text{precisão} \times \text{revocação})}{(\text{precisão} + \text{revocação})} \qquad (7)$$

A seleção e interpretação de qualquer medida para avaliação de desempenho devem ser embasadas nos objetivos do sistema e a tarefa que o usuário realiza (KELLY, 2009). As pesquisas na área de RI têm explorado muitas técnicas, mas os sistemas modernos de recuperação tendem a ser instrumentos fechados, recuperando muitos documentos não relevantes, com elevados erros de precisão, e deixando de recuperar muitos documentos relevantes. Nesse caso, o usuário fica com a importante tarefa de ler ou analisar os resultados obtidos, para determinar se na verdade eles têm as informações solicitadas, e de descobrir como reformular um pedido para novamente verificar se existem documentos relevantes que foram perdidos na busca (WOODS, 2004).

No contexto de emprego de técnicas para indexação e resumo de bases de dados bibliográficas, Lancaster (2003) indica que a avaliação da base não deve ser feita de forma isolada, mas em função de sua capacidade em responder demandas informacionais. O autor cita quatro critérios que devem ser considerados nessa avaliação:

- **Cobertura**: a existência de documentos na base sobre um determinado assunto e de um determinado período;
- **Recuperabilidade**: a capacidade de recuperação de documentos segundo estratégias de busca viáveis;
- **Previsibilidade**: a possibilidade do usuário em aferir a importância dos itens recuperados;
- **Atualidade**: capacidade em fornecer itens novos ou atualizados e não simplesmente entregar elementos antigos.

2.4.5.3 Avaliação em larga escala

Com o objetivo de apoiar a investigação da recuperação de informação do tipo texto, fornecendo infraestrutura para a avaliação em larga escala, o The Text REtrieval Conference (TREC) foi criado em 1992 com frequência anual. Essa conferência, disponível em http://trec.nist.gov/overview.html, é mantida pelo National Institute of Standards and Technology (NIST) e Department of Defense dos EUA. Ela é supervisionada por um comitê de programa composto por representantes do governo, indústria e academia.

Segundo o sítio web do TREC, para cada conferência é fornecido um conjunto grande de teste de documentos e perguntas. Os participantes executam os seus próprios sistemas de recuperação dos dados, e retornam uma lista dos documentos recuperados e classificados. Esses resultados são avaliados e, posteriormente, acontece um fórum para os participantes compartilharem suas experiências. Ainda segundo o sitio, TREC tem cumprido com êxito os seus objetivos em melhorar o estado da arte da RI em texto e de facilitar a transferência de tecnologia. Já existe movimento para recuperação de informação em outras línguas, diferentes do inglês, e de gravações de voz. Além disso, eles afirmam que a maioria dos motores de busca comerciais de hoje incluem alguma tecnologia desenvolvida pela primeira vez em uma conferência TREC.

É interessante destacar que existem outros órgãos que organizam conferências anuais para a realização de avaliação de RI no formato texto, em outras partes do mundo:

- Europa: Conference and Labs of the Evaluation Form (CLEF)

  Sítio web: http://www.clef-initiative.eu/
- Japão: NTCIR (NII Test Collection for IR Systems) Project

  http://research.nii.ac.jp/ntcir/index-en.html
- China: Chinese Web test collection

  http://net.pku.edu.cn/~webg/cwt/en_index.html
- Índia: FIRE (Forum for Information Retrieval Evaluation)

  http://www.isical.ac.in/~clia/

Cada conferência em seu próprio conjunto de *tracks* (trilha temática) que satisfazem necessidades específicas tais como, para a área de saúde, para alguma linguagem específica, para alguma ferramenta midiática etc.

2.4.6 Modelos de recuperação de informação

Baeza-Yates e Ribeiro-Neto (2011) propõem, no Quadro 4, uma definição formal de modelo de RI, que é determinado pelas premissas fundamentais que formam a base de um algoritmo de ranqueamento dos resultados de uma busca.

Quadro 4 – Definição de modelo de RI

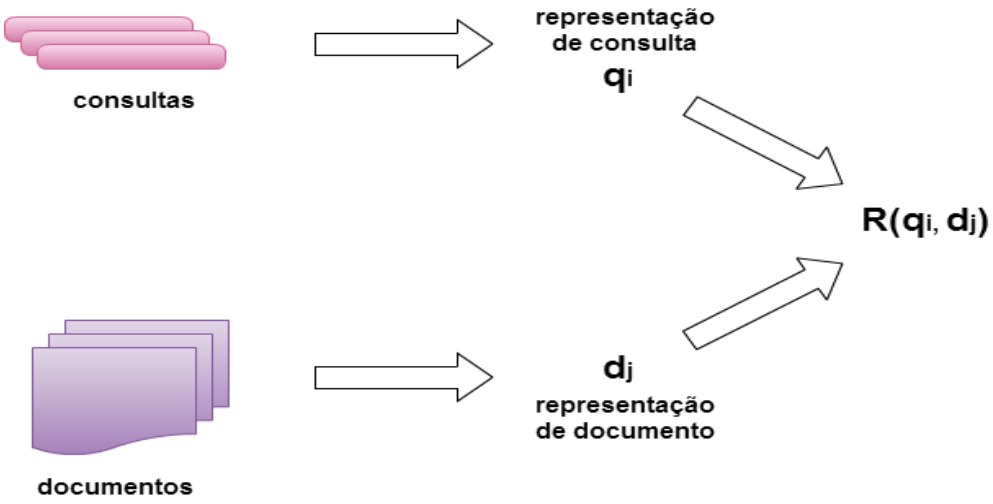
Definição: Um modelo de recuperação de informação é uma quádrupla [**D**, **Q**, ***F***, ***R***(q_i, d_j)] onde:

1. **D** é um conjunto formado por visualizações lógicas (ou representações) da base de documentos.
2. **Q** é um conjunto formado por visualizações lógicas (ou representações) da necessidade informacional do usuário. Tais representações são chamadas de consultas.
3. ***F*** é um *framework* para representações dos documentos do modelo, consultas, e seus relacionamentos, tais como conjuntos e relacionamentos booleanos, vetores e operações de álgebra linear, espaços amostrais e distribuições probabilísticas.
4. ***R***(q_i, d_j) é uma função de ranqueamento que associa um número real com uma representação de consulta $q_i \in$ **Q** e uma representação de documento $d_j \in$ **D**. Tal ranking define uma ordem entre os documentos relacionados à consulta q_i.

Fonte: Baeza-Yates e Ribeiro-Neto (2011, p. 58, tradução nossa)

A função de ranqueamento é uma parte importante nessa definição do Quadro 4, sendo ela que determina um número correspondente ao ranking do documento recuperado para uma dada consulta. Esse fluxo é simbolizado pela Figura 16 onde, uma representação de consulta, $q_i \in$ **Q**, junto com uma representação de documento, $d_j \in$ **D**, retorna um número real por intermédio da função ***R***(q_i, d_j) e de acordo com os processos previsto no framework ***F***.

Figura 16 – Função de ranqueamento $\boldsymbol{R}(q_i, d_j)$



Fonte: Baeza-Yates e Ribeiro-Neto (2011, p. 59, tradução nossa)

Dependendo do modelo de RI adotado haverá diferença entre o resultado do ranqueamento da relevância dos documentos recuperados apresentados ao usuário. Desde os primeiros modelos de recuperação de informação desenvolvidos, chamados de clássicos, tais como os modelos booleano, vetorial e probabilístico, muitos pesquisadores desenvolveram vários outros e inclusive algumas taxonomias para a sua classificação, abordado na próxima subseção, 2.4.7. Compilando as indicações de Salton e McGill (1983), Ferneda (2003), Champclaux, Dkaki e Mothe (2010), Baeza-Yates e Ribeiro-Neto (2011) os três modelos são brevemente caracterizados por:

- **Booleano** (concebido em 1950). É baseado na álgebra de Boole. Documentos são considerados como um conjunto de termos. Uma consulta é formada por uma expressão composta de operadores lógicos, tais como AND, OR e NOT. Um documento é relevante para uma consulta se ele contém os termos da consulta. São vantagens do modelo booleano: facilidade de implementação; e facilidade em expressar a consulta. São desvantagens: a ausência de ordem nos documentos recuperados; como o documento é simplesmente classificado como recuperado ou não, o conjunto resultante por ser vazio ou muito grande.
- **Vetorial** (concebido em 1970). É baseado na teoria de espaço vetorial. Consultas e documentos são considerados como vetores de termos. São atribuídos pesos aos termos das consultas e documentos, que servem para especificar o tamanho e a direção do vetor de representação. O ângulo formado por esses vetores determina a proximidade da ocorrência e também a relevância de cada documento. Salton e McGill

(1983) fizeram estudos aprofundados sobre esse modelo e propuseram o sistema denominado de SMART[41] que foi, posteriormente, amplamente discutido pela literatura.

São vantagens do modelo vetorial: permite restringir o resultado a um número mínimo e máximo de documentos desejados; simplicidade e facilidade para computar similaridades com eficiência; comporta-se bem com base de conhecimento genéricas. São desvantagens: não permite a formulação de consultas booleanas.

- **Probabilístico** (concebido em 1976). Considerado um modelo clássico, é baseado em suposições probabilísticas. Consultas e documentos são considerados como um conjunto de eventos. Um evento representa um termo presente ou ausente. O processo da busca é caracterizado pelo grau de incerteza no julgamento de relevância dos documentos em relação à consulta, sendo a atribuição de relevância uma tarefa do usuário.

São vantagens do modelo probabilístico: baixa complexidade para implementar; garantia de um bom comportamento do método a partir da determinação do princípio probabilístico de ordenação. São desvantagens: não explora a quantidade do termo no documento; ignora o problema da filtragem de informação.

Na literatura encontram-se outros modelos de RI. Alguns desses modelos são resumidos aqui no presente trabalho a partir dos levantamentos realizados por Ferneda (2003), Beppler (2008), Champclaux, Dkaki e Mothe (2010), Baeza-Yates e Ribeiro-Neto (2011):

- **Booleano estendido**. Introduz um peso no termo de consulta do modelo booleano. Usa uma representação de vetor e o cálculo da distância entre os vetores para determinar a relevância de um documento versus uma consulta booleana.
- **Fuzzy**. Baseado na lógica Fuzzy visa superar as limitações do modelo booleano. É baseado em modelagens do mundo real, permitindo trabalhar com a imprecisão, a incerteza, a diversidade, porém, de forma sistemática e rigorosa. Permite introduzir similaridade gradual entre documentos e consultas (valores entre 0 e 1) baseado na teoria de conjuntos.
- **Algoritmos genéticos**. Baseado na teoria da evolução das espécies na biologia. A cada processo de busca, são criadas novas estruturas, que são selecionadas baseadas no conceito de *fitness* (adequação às condições ambientais). Depende diretamente das

[41] Projeto SMART (System for the Manipulation and Retrieval of Text) iniciou-se em 1961 na Universidade de Harvard.

interações do usuário, para que a base de conhecimento fique cada vez mais apta a dar resultados mais relevantes, para uma dada situação problema da RI.

- **Redes neurais**. Tenta simular computacionalmente o funcionamento dos neurônios cerebrais, onde as expressões de busca do usuário e os documentos relacionam-se com os termos de indexação, comportando-se como uma rede neural. O resultado de uma busca é formado por documentos ligados à expressão de busca e também por documentos que o sistema inferiu a partir da relevância registrada em interações anteriores.
- **Indexação semântica latente**. Tenta trabalhar com o lado semântico de cada termo do documento, introduzindo uma interessante conceituação de problema de RI baseado na decomposição em valores singulares. Mapeia cada vetor de documento e consulta em um espaço dimensional composto de conceitos. Do ponto de vista prático, ainda não existem resultados satisfatórios.
- **Best Match 25**. Foi criado a partir de uma série de experimentos e variações sobre o modelo probabilístico clássico. Porém, diferentemente do modelo clássico, é possível calcular o *ranking* dos documentos recuperados sem informações de relevância fornecidas pelo usuário. Também usado para avaliar novos métodos de ranqueamento em substituição ao modelo vetorial.
- **Processamento de linguagem natural**. É usado em sistemas com processamento de linguagens naturais[42]. Considerado um modelo semântico, usa a estrutura e o significado dos documentos definindo um modelo de linguagem para cada documento e usa esse modelo para investigar a probabilidade de gerar uma dada consulta para ele. O *ranking* de relevância é produzido por intermédio da ordenação dessas probabilidades.
- **Divergência de aleatoriedade**. Com características do modelo de linguagens, esse modelo computa o peso dos termos, pela medida da divergência entre uma distribuição de termo produzido por um processo randômico e a real distribuição de termos. Nem todas as palavras são importantes para descrever o conteúdo de um documento então elas são distribuídas aleatoriamente.
- **Redes bayesianas**. É uma abordagem para o desenvolvimento do modelo probabilístico. Redes bayesianas são grafos direcionados e não cíclicos, onde os nós representam variáveis e as arestas representam as probabilidades de relacionamento

[42] Linguagens naturais são usadas para a comunicação humana. Exemplos: Português, Inglês, Espanhol etc.

entre elas. Dessa forma, as consultas e os documentos formam uma rede que capaz de representar interdependências entre termos e documentos.

- **Inferencial**. Baseado nas redes bayesianas, esse modelo é considerado uma visão epistemológica do problema de recuperação de informação. A rede construída é formada por nós que representam documentos, termos de índice, consultas e necessidades informacionais do usuário. As arestas representam a confiança entre esses elementos. De uma maneira geral, a relevância de um documento para uma consulta corresponde ao grau de confiança com que o documento satisfaz à necessidade do usuário.
- **Texto estruturado**. Esse modelo aproveita-se da estrutura existente em textos para melhorar vários estágios do processo de RI: indexação, recuperação, apresentação dos resultados e consulta. Usa linguagens de marcação, tal como XML.
- **Web**. São modelos que trabalham sobre a web. Considerando as características bem peculiares da web (grande, volátil e distribuído com informações não estruturadas). O ranqueamento é a parte mais complexa, porém importante, do processo de busca e geralmente é realizada em função dos *links* associados ao documento. Além disso, a identificação da qualidade da informação recuperada é um dos grandes desafios.
- **Grafo**. Explora a estrutura de grafo dos documentos para determinar similaridade entre uma consulta e um documento: a similaridade não depende somente dos termos indexados e compartilhados, mas também de sua vizinhança.
- **Multimídia**. São modelos que trabalham na recuperação integrada de texto, imagem, vídeo e som. É uma RI complexa, diferentemente do texto que possui palavras que se comportam como unidades básicas de leitura, os elementos multimídia possuem poucos delimitadores. A distância semântica e complexidade aumentam na seguinte direção: texto, fala, imagem, vídeo e música. A determinação do grau de similaridade entre os elementos multimidiáticos é determinante para melhorar a probabilidade de o usuário achar respostas relevantes.

2.4.7 Taxonomias para recuperação de informação

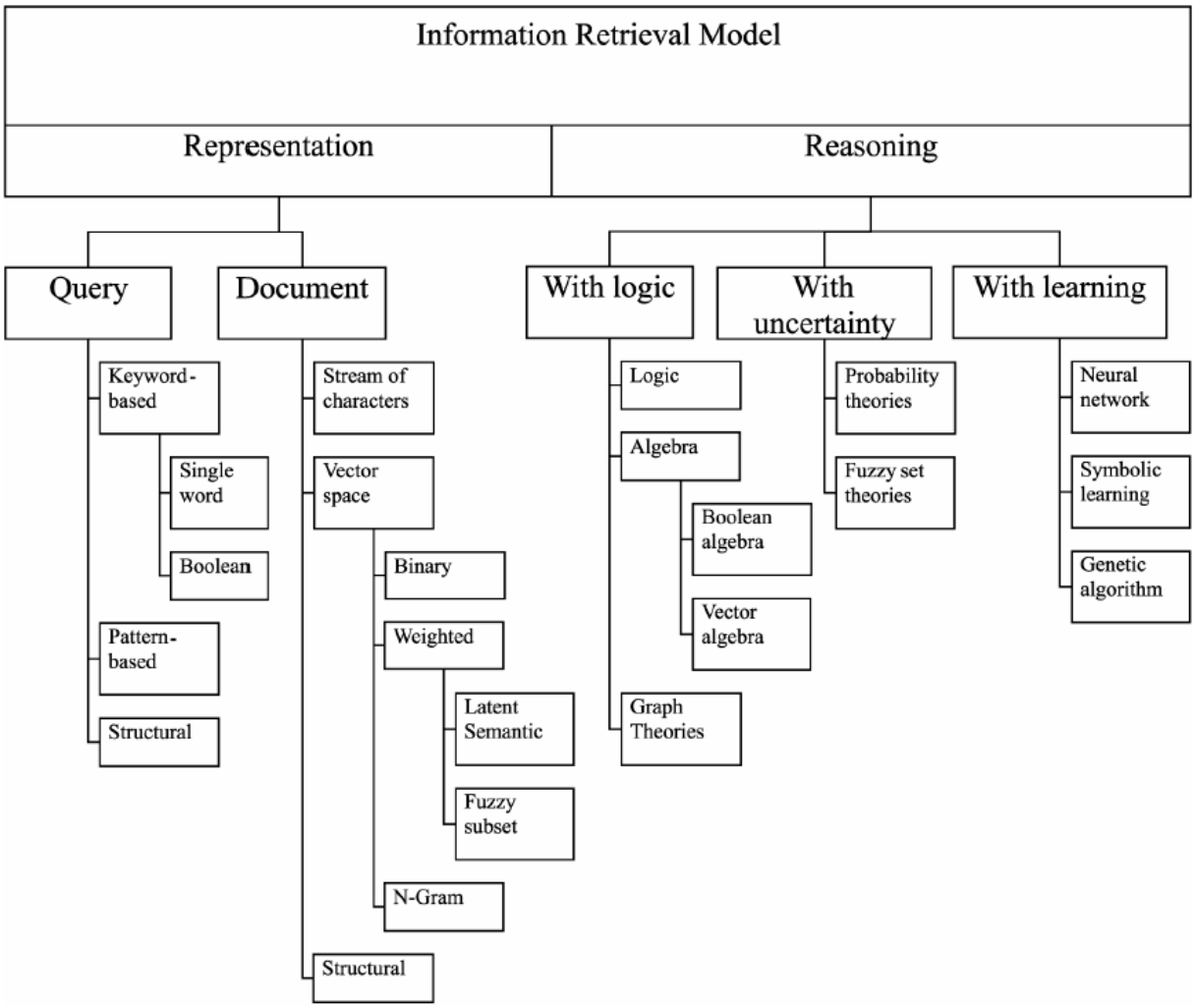
Existe uma proliferação de algoritmos, métodos, tecnologias e ferramentas para RI além das diferentes terminologias usadas (CANFORA; CERULO, 2004). As taxonomias são úteis para organizar esse conhecimento, facilitar o estudo do estado da arte, bem como

auxiliar a pesquisa para o desenvolvimento de novos modelos com novas características para atender melhor às necessidades informacionais. Diversas taxonomias para classificar modelos de RI são encontradas na literatura. Kuropka (2004) desenvolveu uma taxonomia com duas vertentes: a primeira classifica os modelos segundo a sua base matemática, ou seja, se são fundamentados na teoria de conjuntos, se são algébricos ou se são probabilísticos. A segunda vertente é relacionada à existência de dependências entre os termos da pesquisa. Champclaux, Dkaki e Mothe (2010), propuseram uma taxonomia que faz analogia com a Ciência Cognitiva usando critérios de similaridade para comparar um objeto com outro. Esses autores classificam os modelos segundo três abordagens de similaridade: (i) distância espacial, (ii) baseado em atributos e (iii) estrutural. A taxonomia de Naranjo, Kauffman e Ferrández (2014) é baseada no nível de incerteza, sendo classificada em probabilístico, grau de possibilidade e o quanto é relacionado a fatos. Outras duas taxonomias, Canfora e Cerulo (2004) e Baeza-Yates e Ribeiro-Neto (2011), que se aproximam mais das características do presente trabalho, são apresentadas com mais detalhes a seguir.

Canfora e Cerulo (2004) propõem uma taxonomia para modelos de RI, Figura 17 e Figura 18, em duas visões, vertical e horizontal:

- **Vertical**: representado pela Figura 17, classifica os modelos segundo um conjunto de características básicas, ou seja, o tipo de representação (*representation*) da consulta (*query*) e do documento (*document*) e também o método de condução da recuperação (*reasoning*) que pode ser com lógica (*with logic*), com incerteza (*with uncertainty*) e com aprendizagem (*with learning*). Espera-se que um dado modelo seja enquadrado em um elemento do ramo *Query*, em outro do ramo *Document* e em, pelo menos, mais um do ramo *Reasoning*.

Figura 17 – Taxonomia vertical de modelos de RI

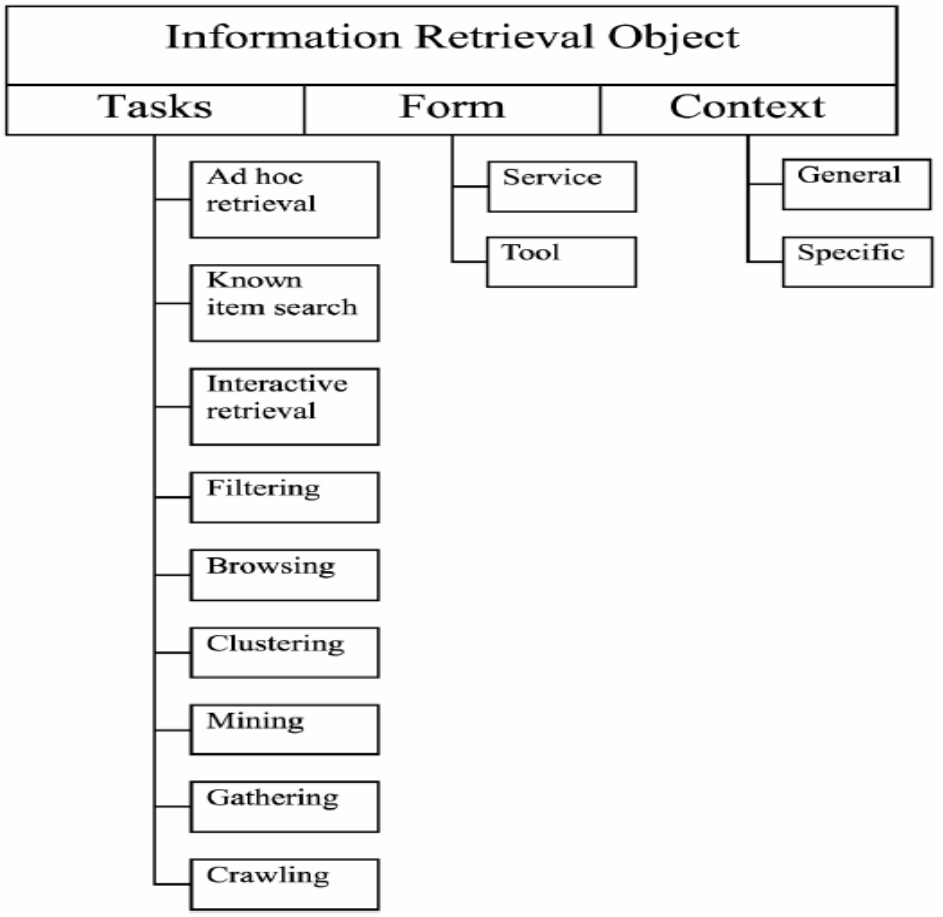


Fonte: Canfora e Cerulo (2004, p. 177)

- **Horizontal**: representado pela Figura 18, classifica os objetos da RI, ou seja, artefatos que resolvem problemas da RI, e são identificados por três componentes: tarefas (*Tasks*), forma (*Form*) e contexto (*Context*). Existem 9 categorias de tarefas, não mutualmente exclusivas: (i) *ad hoc retrieval*, quando ocorrem consultas independentes sem depender da continuidade de interatividade com o usuário; (ii) *known item search*, tal como ocorre com *ad hoc retrieval*, porém o resultado da pesquisa é conhecido pelo usuário; (iii) *interactive retrieval*, onde o julgamento do usuário sobre informações já recuperadas é determinante ainda durante o processo da recuperação; (iv) *filtering*, onde cada documento é categorizado em classes, sendo estas escolhidas a priori pelos usuários para que suas buscas já sejam automaticamente filtradas por elas; (v) *browsing*, quando o usuário quer explorar a base de conhecimento simplesmente navegando-a, sem especificar uma busca; (vi) *clustering*,

quando usa um reconhecimento automático, por intermédio de alguma medida de similaridade, para agrupar documentos em categorias e assim melhorar o processo de RI; (vii) *mining*, quando usa um processo automático de extração de informações chaves dos documentos; (viii) *gathering*, que tem a capacidade de realizar a recuperação de informação em fontes heterogêneas de informações, tal como ocorre nas máquinas de meta-busca quando apresentam único resultado advindo de várias máquinas de busca; (ix) *crawling*, que se concentra na atualização constante da fontes de informação que normalmente são processadas por sucessivas atividades de busca. O ramo denominado ‘forma’ representa como o objeto é entregue ao usuário e pode ser de serviço (*service*), onde há a entrega do serviço que irá proporcionar a busca tal como acontecem com as máquinas de busca na web; ou de ferramenta (*tool*), quando um software é instalado no cliente. Quanto ao ramo ‘contexto’, ele pode ser geral (*general*), onde trabalha num domínio de conhecimento mais amplo; ou específico (*specific*), quando trabalha num domínio específico do conhecimento.

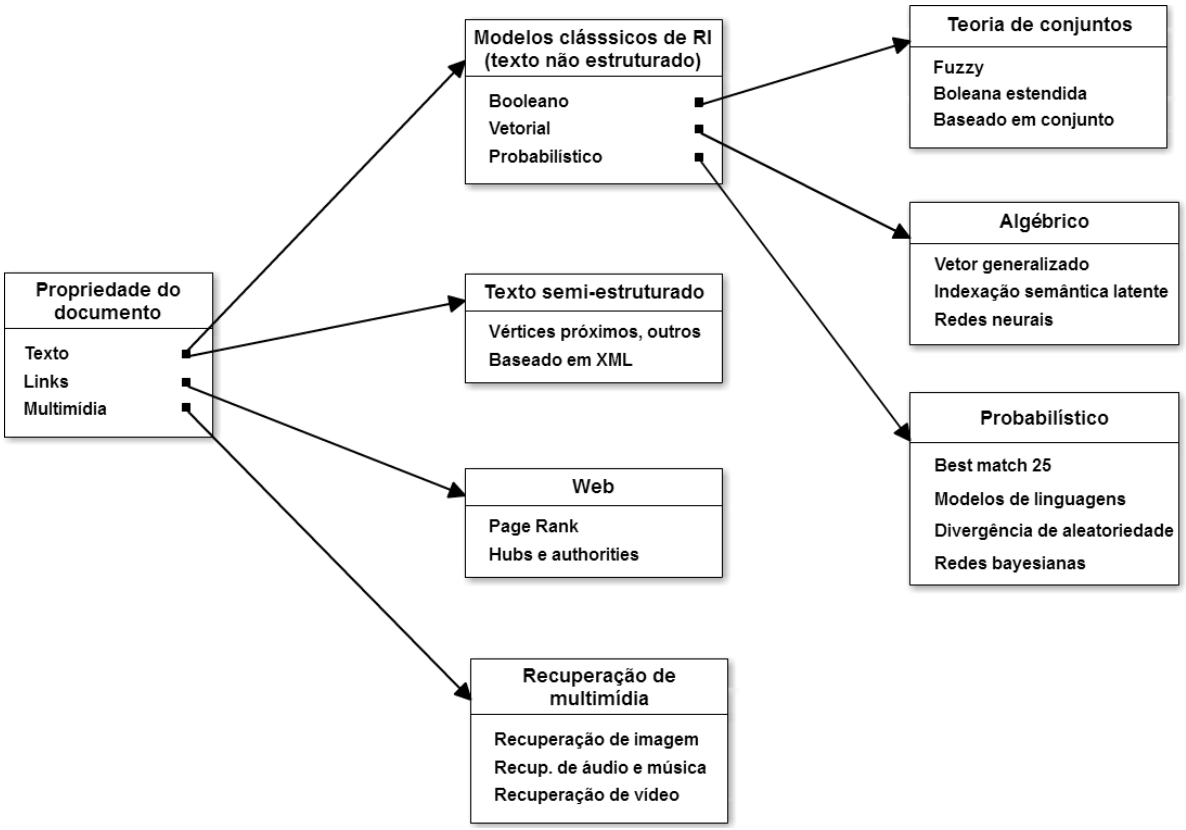
Figura 18 – Taxonomia horizontal de modelos de RI



Fonte: Canfora e Cerulo (2004, p. 184)

A taxonomia proposta por Baeza-Yates e Ribeiro-Neto (2011), representada na Figura 19, distingue três principais tipos de modelos de RI: aqueles baseados em texto, os baseados em *links* e os baseados em objetos multimídia. Modelos baseados em texto podem ser estruturados ou não, sendo que os não estruturados classificam-se em um dos três modelos clássicos (booleano, vetorial e probabilístico), ainda com suas derivações e especializações. Para modelos baseados em *links* os autores criaram uma classificação própria denominada Web. Finalmente, modelos baseados em objetos multimídia classificam aqueles que trabalham com imagem, música e vídeo.

Figura 19 – Taxonomia geral de modelos de RI



Fonte: Baeza-Yates e Ribeiro-Neto (2011, p. 60, tradução nossa)

A taxonomia proposta por Canfora e Cerulo é ampla em seus critérios e consegue classificar os modelos utilizando-se vários elementos e suas especificidades, porém ela é específica para recuperação de informação no formato de texto, enquanto Baeza-Yates e Ribeiro-Neto, apesar de proporem uma taxonomia mais simples em se tratando de critérios, conseguiram ampliar o espectro de tipos de documentos, tratando também os documentos formados por *links* e aqueles do tipo multimídia, além do texto.

2.4.8 Considerações finais da seção

Observa-se que a área da RI cognitiva (subseção 2.4.2.3) faz uma boa interseção com a área da RI interativa (subseção 2.4.2.1), pois ambas se preocupam com a relação entre o usuário e o sistema de RI, e sugerem métodos para melhorá-la. Percebe-se aproximação ainda maior da RI cognitiva com a área comportamento informacional (subseção 2.4.2.2), especificamente *information seeking behavior*, por ambas serem dependentes dos estados cognitivos do usuário. Porém, *information seeking behavior* ainda é mais ampla do que a RI cognitiva por cuidar dos estados afetivos do usuário, ambientes sociais, culturais e organizacionais.

Apesar da existência da RI cognitiva e do comportamento informacional, que tratam o usuário como elemento fundamental no processo de RI, a tendência é que todos os modelos de RI aumentem cada vez mais a preocupação com o usuário. Além disso, há também uma tendência para que os processos de RI sejam menos estáticos e mais interativos como a *interactive information retrieval* discutida na subseção 2.4.2.1. Porém, ainda há muito o que fazer nos sistemas de RI, antes de considera-los plenamente interativos. Existem lacunas quanto ao processo interno de escolha da informação recuperada relevante, para estabelecer um bom ranqueamento e oferecer ao usuário documentos mais sintonizados à sua necessidade informacional. A formulação de modelos híbridos de RI, que aproveitam-se de características boas dos modelos existentes e bem amadurecidos, pode ser um caminho promissor.

Outra proximidade que pode ser observada é entre a área de recuperação de conhecimento (subseção 2.4.3) com as áreas visualização de informação e visualização de conhecimento (subseção 2.4.4). Apesar da recuperação de conhecimento cuidar de várias etapas anteriores, em ambos os casos a apresentação da informação ocorre por meio de estruturas com o objetivo de melhorar a criação do conhecimento e trazer novas percepções aos usuários.

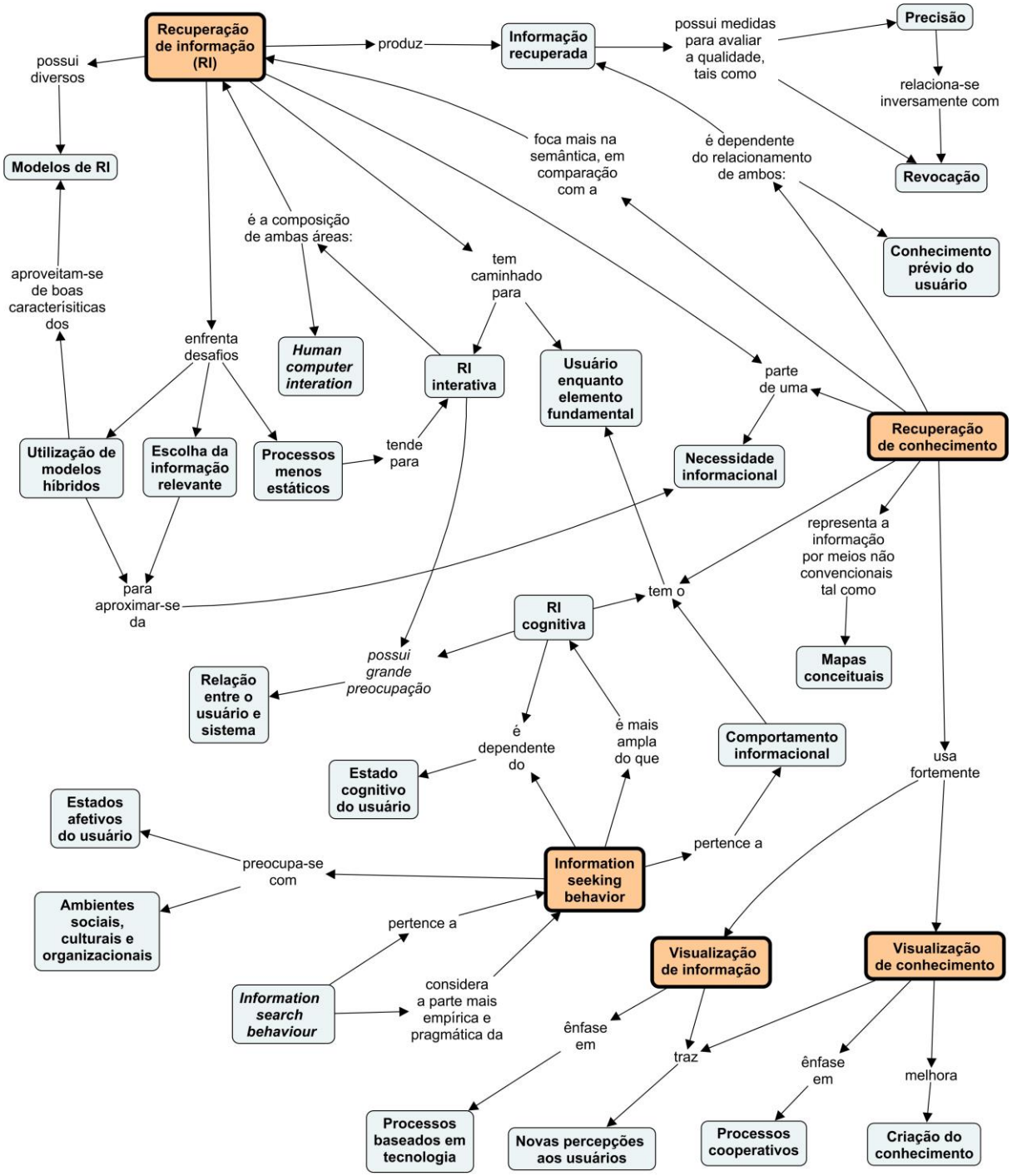
Outra vertente são os avanços da RI que se aproveitam de conhecimentos de áreas emergentes. Por exemplo, Gurrin *et al.* (2010) citam o exemplo de um sistema que recupera um link quebrado de um site substituindo-o pela página correta através de técnicas de mineração de textos, Araújo Junior (2007, p. 149) defende a tese de que a "[...] mineração de textos pode, em associação com o processo de indexação manual, trazer ganhos no índice de precisão no processo de busca e recuperação da informação [...]" (p. 149). Duque (2005) explora o uso da linguística computacional associada a ontologias para otimizar o desempenho de sistemas de RI por meio da utilização de técnicas que permitam contextualizar

as palavras dos textos a serem indexados. Beppler (2008) propôs um ambiente de busca interativo, baseado na hermenêutica, onde o usuário navega em conceitos de uma ontologia de domínio usada para a construção de indexadores da base de conhecimento.

Yao *et al.* (2007), após a análise de vários artigos sobre recuperação de conhecimento, afirmam que essa área é o próximo passo da RI e sugerem que é preciso estudá-la de uma forma mais intensa e superar muitos desafios, por exemplo, o de lidar com o grande volume de dados da web, já que, tradicionalmente, bases de conhecimento são armazenadas localmente. A recuperação de conhecimento aliada à visualização de conhecimento podem mehorar o relacionamento entre os sistemas de RI e a necessidade informacional do usuário.

O mapa conceitual da Figura 20 apresenta alguns relacionamentos importantes abordados nessa seção sobre recuperação de informação e conhecimento, destacando, em cor alaranjada e espessura maior, alguns conceitos relevantes para a presente tese. Entre as várias proposições existentes no mapa, destacam-se os desafios da RI e as medidas para avaliação da qualidade da informação recuperada: precisão e revocação. Também se destaca o usuário enquanto elemento fundamental da RI cognitiva, do comportamento informacional, da recuperação de conhecimento, e como tendência a ser seguida pela RI. Outras proposições cobrem as áreas de visualização de informação e conhecimento e suas ênfases. A área *information seeking behaviour*, pertencente à área comportamento informacional, aparece bastante abrangente envolvendo elementos afetivos, sócio culturais e cognitivos do usuário. Observa-se também a necessidade informacional como um elemento fundamental no contexto e que faz movimentar os processos de recuperação tanto de informação quanto de conhecimento.

Figura 20 – Mapa conceitual com alguns relacionamentos abordados na seção 4: recuperação de informação e conhecimento



Fonte: Elaboração própria

2.5 Ciência das Redes

O trabalho em redes é tão antigo quanto à história da humanidade, mas, somente nas últimas décadas, as pessoas passaram a observá-lo como uma ferramenta organizacional (MARTELETO, 2001). Apesar das diferenças aparentes, muitas redes emergem e evoluem, baseadas em um conjunto fundamental de leis e mecanismos, e estes são a base dessa nova ciência, chamada de Ciência das Redes (BARABÁSI, 2013). Essa ciência representa os estudos científicos realizados em redes, e teve o seu desenvolvimento impulsionado pela tecnologia (NATIONAL RESEARCH COUNCIL, 2005). Ciência das redes é uma área interdisciplinar que estuda teoria e métodos sobre as redes complexas. O seu aspecto interdisciplinar pode ser constatado em trabalhos publicados em áreas como: Ciência da Informação, Ciências Sociais, Ciências Humanas, Sociologia, Matemática, Estatística, Física, Ciência da Computação etc. É interessante também observar como sociólogos, físicos, biólogos e outros cientistas encontraram inesperadamente conexões entre o funcionamento do mundo humano e a mecânica de outras coisas aparentemente muito diferentes: da célula viva e o ecossistema global à internet e o cérebro humano (BUCHANAN, 2002).

Entre os vários acontecimentos históricos e importantes da Ciência das Redes, dois são destacados aqui como marcos importantes do nascimento e estudo sistematizado das redes. O primeiro deles é atribuído a Leonard Euler, matemático suíço, que em 1736 iniciou uma área de estudos na Matemática, hoje conhecida por Teoria dos Grafos[43], quando escreveu a respeito de um conhecido problema da época sobre as sete pontes da cidade de Königsberg[44]. O desafio era comprovar analiticamente como seria possível passar pelas sete pontes sem repetir nenhuma delas. Euler fez fama quando provou, de forma simples e elegante, que o problema não tinha solução (BARABÁSI, 2002). O segundo marco foi uma contribuição de Jacob Levy Moreno, psiquiatra romeno que iniciou a área de estudo denominada Sociometria[45], quando desenvolveu pesquisas para analisar pequenos grupos de pessoas, ao invés da sociedade inteira, questionando-as sobre suas escolhas sociais, como por exemplo: ‘quem você escolheria como um amigo/colega/consultor?’ Essas escolhas sociais são consideradas, na Sociometria, a maior expressão das relações sociais e são representadas por meio de grafos (NOOY; MRVAR; BATAGELJ, 2011). Depois desses dois marcos na história

[43] Teoria dos Grafos é um ramo da Matemática que estuda grafos - é um conjunto de vértices e um conjunto de linhas entre quaisquer pares de vértices (NOOY; MRVAR; BATAGELJ, 2011). A Teoria dos Grafos fornece base Matemática para a Ciência das Redes (BARABÁSI, 2014).
[44] Königsberg: cidade pertencente à Prússia no século XVIII, atualmente Kaliningrado pertencente à Rússia.
[45] Sociometria estuda as relações interpessoais e considera que a sociedade é formada por grupos inter-relacionados (NOOY; MRVAR; BATAGELJ, 2011).

da Ciência das Redes, muita pesquisa foi desenvolvida estabelecendo um pleno crescimento até os dias de hoje, principalmente na aplicabilidade do desenvolvimento de soluções para problemas na sociedade.

Essa seção aborda algumas importantes propriedades e fenômenos selecionados sobre as redes complexas. Aborda também a análise de redes complexas e algumas métricas e medidas selecionadas. As seleções feitas, dentro de um vasto e amplo escopo da área Ciência das Redes, foi realizada para atender especificamente o desenvolvimento da tese. A primeira subseção apresenta uma visão geral de redes complexas. A segunda subseção discute algumas propriedades e fenômenos de redes complexas. A terceira subseção discute a análise de redes complexas. Finalmente, a quarta subseção trata de algumas medidas e métricas de rede.

2.5.1 Redes complexas

Para fins desse trabalho, uma rede é uma coleção de pontos, chamados também de nós ou vértices, juntos em pares através de linhas, chamadas de ligações ou arestas (NEWMAN, 2010). As redes podem ser representadas por um grafo, no qual informações adicionais são colocadas nos vértices e linhas que, apesar de importantes, são irrelevantes para determinar a estrutura da rede, porque esta depende apenas do padrão de suas ligações (NOOY; MRVAR; BATAGELJ, 2011). A nomenclatura desses elementos, como descreve Adamic (2013) no Quadro 5, varia conforme a área de conhecimento. O presente trabalho adota, na maior parte das vezes, os termos nó e ligação/conexão. Outro conceito básico em uma rede é o grau de um nó que é igual a quantidade de ligações que ele tem com outros nós da rede.

Quadro 5 – Nomenclatura encontrada na literatura inglesa sobre elementos de rede

Points (Pontos)	*Lines* (Linhas)	*Field* (Área do conhecimento)
vertex (vértices)	*edges*, *arcs* (arestas e arcos)	*Math* (Matemática)
nodes (nós)	*links* (ligações)	*Computer Science* (Ciência da Computação)
sites (sítios)	*bonds* (amarrações)	*Physics* (Física)
actors (atores)	*ties*, *relations* (vínculos, relacionamentos)	*Sociology* (Sociologia)

Fonte: Adaptado de Adamic (2013)

Uma rede complexa descreve uma ampla variedade de sistemas na natureza e na sociedade (ALBERT; BARABÁSI, 2002), sendo um grafo normalmente modelado a partir de sistemas reais, que apresenta características topológicas não triviais. O estudo das redes complexas surgiu a partir de desdobramentos da teoria dos grafos. Segundo Newman (2010), as redes complexas mais comuns são as redes tecnológicas, as redes de conhecimento, as redes biológicas e as redes sociais. Porém, de uma maneira geral, em várias áreas do conhecimento humano existem muitos sistemas de interesse para os cientistas que são compostos de peças individuais ou componentes, ligados entre si de alguma forma. Além disso, não se deve confundir a análise de redes complexas com o estudo de sítios e aplicações de mídias sociais tais como Facebook[46], Whatsapp[47] e Twitter[48], pois a primeira estuda os padrões de organização de quaisquer coletivos complexos, em quaisquer ambientes, humanos ou não, enquanto que a segunda estuda a troca de informações em comunidades virtuais humanas (FRANCO, 2012).

2.5.2 Propriedades e fenômenos em redes complexas

Nas próximas subseções são apresentados alguns tipos, princípios, propriedades ou fenômenos importantes que ocorrem em redes complexas. Esses elementos não são obrigatórios nem disjuntos, isto é, uma rede complexa pode apresentar alguns ou vários deles. São também fornecidos alguns exemplos de redes específicas nas quais eles ocorrem.

2.5.2.1 Rede de informação

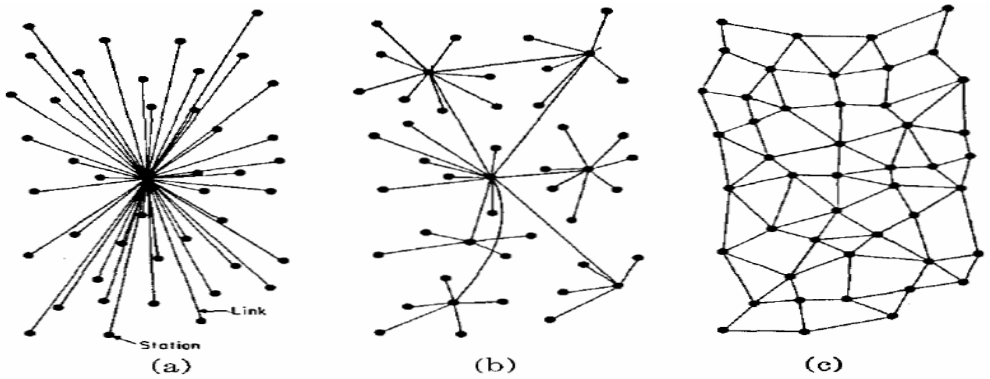
As redes de informação são redes constituídas por itens de informação ligados. Elas são, na maioria das vezes, redes complexas e tem como exemplos típicos a web e as redes de citações bibliográficas entre autores. Além disso, existem algumas redes que podem ser consideradas redes de informação, mas que também têm aspectos sociais, como as redes de comunicação de email, redes em sítios web e aplicações de mídias sociais e revistas on-line (NEWMAN, 2010). Newman ainda alerta que não pode ser disjunta a classificação de redes em sociais, de informação entre outros tipos, havendo muito exemplos que se encaixam em vários tipos. Um mapa conceitual é também um exemplo de rede de informação, pois liga conceitos por intermédio de frases de ligação, formando as proposições.

[46] Facebook, disponível em <https://www.facebook.com/>.
[47] Whatsapp, disponível em <https://www.whatsapp.com/>.
[48] Twitter, disponível em <https://twitter.com/>.

2.5.2.2 Rede centralizada, descentralizada e distribuída

Motivado pelo desenvolvimento de um sistema de comunicação que sobrevivesse a um ataque nuclear, Paul Baran, em 1964, elaborou a topologia de rede distribuída, onde cada nó é conectado a todos os adjacentes, sem a ocorrência de centros (BARABÁSI, 2002). A Figura 21 mostra a diferença entre três conhecidas topologias sugeridas por Baran (1964): (a) centralizada, (b) descentralizada e (c) distribuída. Franco (2012) chama atenção para o fato de que, contrário à crença popular, descentralizado não é sem centro, mas, com muitos centros.

Figura 21 – Diferença entre rede centralizada, decentralizada e distribuída



Fonte: Baran (1964)

Em seu artigo, Baran (1964), demonstra a vantagem de uma configuração distribuída, em termos de capacidade de sobrevivência em casos de ataque inimigo contra nós da rede, *links* ou combinações de nós e *links*, em comparação a outras topologias de rede.

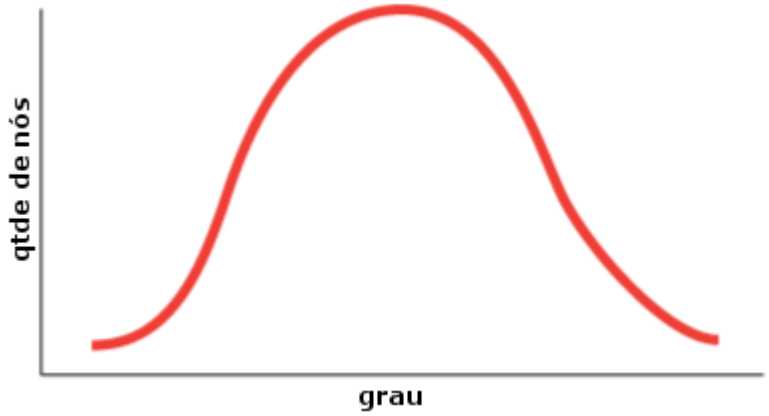
2.5.2.3 Rede aleatória

Erdös e Rényi (1959) influenciaram muito o estudo das redes aleatórias com sua pesquisa, sobre a distribuição Poisson[49] de frequência dos graus dos nós de uma rede, onde a maioria dos nós tem um grau próximo da média. O Gráfico 4 representa de forma aproximada a distribuição dos nós em função do seu grau, ou seja, existe maior quantidade de nós com a faixa de grau médio e poucos nós nos dois extremos, com poucas e muitas conexões. Redes

[49] Distribuição de Poisson é uma distribuição de probabilidade de variável aleatória discreta que expressa a probabilidade de eventos ocorrerem num período de tempo e de forma independente da ocorrência do último evento.

com esse tipo de distribuição de frequência tendem a não gerar aglomeração de nós e nem *hubs*[50]. O Gráfico 4 também é conhecido como curva do sino.

Gráfico 4 – Distribuição de frequência de graus de nós de uma rede segundo a distribuição de Poisson aproximada

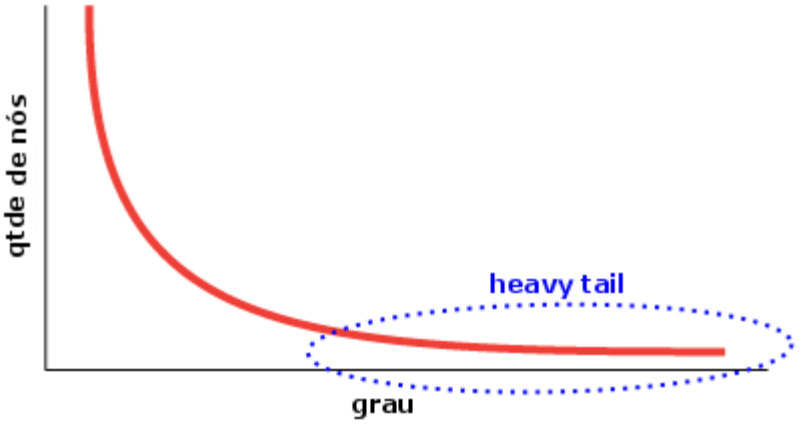


Fonte: Elaboração própria

2.5.2.4 Rede livre de escala ou scale-free

Uma rede livre de escala, ou *scale-free*, é uma rede na qual a distribuição de frequência dos graus obedece à lei de potência (*power law*), como mostra o Gráfico 5.

Gráfico 5 – Distribuição de frequência de graus de nós de uma rede segundo a lei de potência



Fonte: Elaboração própria

[50] Hub: no contexto da Ciência da Redes, hub é um nó com uma quantidade de conexões bem maior que a média.

Em redes desse tipo, não se pode escolher um nó com grau ‘típico’, e por essa razão tais redes são chamadas de livre de escala, ou *scale-free*. Uma propriedade importante das redes desse tipo é a sua resistência contra eliminações aleatórias, pois se alguns nós são removidos de forma aleatória a rede ainda permanece conectada em sua grande parte. Por conta dessa capacidade de ‘sobreviver’, Barabasi (2002) chamou as redes de tais tipos, metaforicamente, de ‘teia sem aranha’ (*web without a spider*).

Até finais dos anos 90, a web era considerada uma rede aleatória. Porém, com o seu crescimento nos anos seguintes, experimentos de Albert, Jeong e Barabási (1999) mostraram que ela não tinha a conhecida estrutura de rede aleatória, com distribuição de frequência de graus em forma de sino do Gráfico 4, bastante estudada por Erdös e Rényi (1959). Albert, Jeong e Barabási (1999) demonstraram que a distribuição de frequência dos graus dos nós da web obedece à lei de potência, e portanto apresenta a forma mostrada no Gráfico 5. A distribuição de frequência dos graus dos nós da web decai muito menos rapidamente do que uma curva de Poisson, e ao invés de praticamente todos os nós terem mais ou menos o mesmo grau, espera-se que alguns sejam altamente conectados e a grande maioria tenha um grau menor do que a média. Além da web, outras redes também são consideradas livre de escala, por exemplo, a internet e a rede metabólica de organismos, onde cada nó é um metabolito e as ligações são as reações entre eles (BARABÁSI, 2013).

As redes livre de escala tem uma propriedade conhecida por *heavy tail*, como mostra o destaque no Gráfico 5, que numa tradução literal tem o significado ‘rabo pesado’ ou ‘rabo gordo’. Isso porque poucos nós possuem grau elevado deixando o ‘rabo’ da curva mais ‘pesado’, enquanto que a grande maioria possui grau pequeno. Na web esse fenômeno é representado pelos poucos sites que possuem uma quantidade muito grande de conexões, em relação à maioria restante.

2.5.2.5 Rede mundo pequeno ou small-world

O fenômeno mundo pequeno, ou *small-world*, ocorre numa rede quando dois nós quaisquer estão susceptíveis a estarem ligados por um caminho relativamente curto. Esse princípio também é conhecido pelo termo cunhado por Jonh Guare[51] como *six degrees of separation* (os seis graus de separação), originado do estudo feito por Milgram (1967). Devido a essa ligação curta entre os nós, ela possui baixo diâmetro (subseção 2.5.4.2). Além

[51] John Guare: escritor irlandês e americano, escreveu a famosa peça teatral *Six Degrees of Separation*, premiada em 1990, e transformada em filme em 1993.

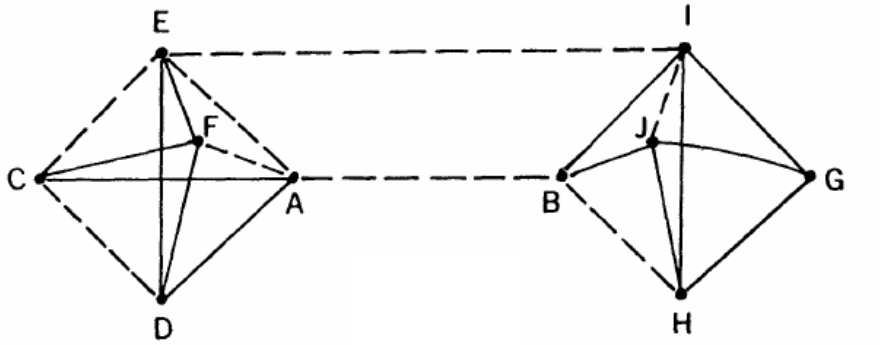
disso, ela é uma rede com alto coeficiente de clusterização (subseção 2.5.4.1), isto é, a probabilidade de dois nós conectados entre si terem um ou mais vizinhos em comum é grande.

Apesar da ideia dos mundos pequenos ser notavelmente simples, ela aparece em uma quantidade muito grande de estruturas do mundo real, desde o cérebro humano, passando pelas redes que unem a sociedade até as relações entre os termos nas línguas que usamos para falar e pensar (BUCHANAN, 2002). A web é um exemplo de rede que também tem a propriedade mundo pequeno (BARABÁSI, 2013). Albert, Jeong e Barabási (1999) mediram e acharam uma média de 19 conexões entre dois nós quaisquer da web. Watts e Strogatz (1998) citam redes que satisfazem o princípio *small world*, como a rede de energia do oeste dos EUA, a rede de colaboração entre os atores de cinema e redes de propagação de doenças.

2.5.2.6 A força dos laços fracos ou the strength of weak ties

De forma pioneira, o sociólogo Mark Granovetter (1973), descreve o termo ‘a força dos laços fracos’ ou ‘*the strength of weak ties*’. Ele explica que a ligação de uma pessoa em uma rede social pode ser dividida em laços forte e laços fracos. Os laços fortes ligam pessoas próximas, com as mesmas crenças, costumes, visões e valores, e pertencem a grupos com grande quantidade de conexões, como as ligações mostradas na rede da Figura 22 nos dois grupos de nós A-C-D-E-F e B-G-H-I-J. Os laços fracos ligam pessoas mais distantes, sem intensidade de relacionamento. As conexões pertencentes às fronteiras dos dois grupos, como as ligações E-I e A-B da mesma figura, tendem a ser laços fracos, porém elas conseguem atravessar fronteiras e são cruciais na comunicação de informações importantes para a sua vida, com possibilidades de conduzir informações sobre grandes oportunidades.

Figura 22 – A força dos laços fracos de Granovetter



Fonte: Granovetter (1973)

Em outro artigo, Granovetter (1983) destaca que, apesar dos laços fracos oferecerem acesso à informação e recursos que vão além daquelas que estão disponíveis no círculo social de um indivíduo, os laços fortes são, obviamente, também importantes pela sua facilidade de disponibilidade, desempenhando papel único.

Onnela *et al.* (2007) examinaram os padrões de comunicação de milhões de usuários de telefones celulares em função das conversas entre eles e, de uma maneira geral, constataram a presença do fenômeno ‘a força dos laços fracos de Granovetter’. Segundo Kleinberg (2013), enquanto há um notável nível de complexidade nas interações de uma rede social de relações de trabalho, por outro lado existe expectativa considerável na utilização de grande volume de dados extraídos da web, representando grandes populações, combinadas com modelos matemáticos e computacionais para estudar os fenômenos implicados nos laços fracos.

2.5.2.7 Ligação preferencial ou preferential attachment

A ligação preferencial, ou *preferential attachment*, é um fenômeno que decorre do crescimento de uma rede complexa, caracterizado pelo fato de novos nós da rede estarem mais susceptíveis a se vincularem a nós já bem estabelecidos, isto é, a nós com maior grau. Por exemplo, é comum as redes sociais terem poucos nós com elevado grau e muitos nós com grau baixo e, assim, a fixação preferencial ocorre quando novos membros preferem se ligar àqueles com maior grau. Esse fenômeno concretiza alguns ditos populares, como ‘os ricos ficam mais ricos’ ou ‘o sucesso gera sucesso’ (NOOY; MRVAR; BATAGELJ, 2011). Assim, graças a esse fenômeno, quem já é popular numa rede social tende a ficar cada vez mais popular com o seu crescimento.

Hidalgo *et al.* (2007) verificaram que a rede de relacionamentos entre os produtos exportados entre países tende a concentrar a maioria dos produtos de luxo num núcleo densamente conectado, enquanto os produtos de menor renda ocupam uma periferia menos conectada entre os países, representando assim a propriedade de ligação preferencial. A web é também um exemplo de rede que exibe o fenômeno de ligação preferencial (BARABÁSI; ALBERT, 1999). Há uma incidência de poucos sítios web com uma quantidade muito grande de conexões e a tendência deles aumentarem cada vez mais.

2.5.2.8 Aptidão ou fitness

Bianconi e Barabási (2001 apud BARABÁSI, 2013) observaram que a ligação preferencial não era o único fator que explicava a distribuição de graus dos nós na web, pois alguns poucos sítios da web se tornam grandes *hubs* muito rapidamente e não somente pelo fato deles já estarem bem estabelecidos. Esse novo fenômeno foi descrito pelos autores como aptidão (*fitness*), ou seja, o crescimento dos graus ocorre ao longo do tempo, mas a sua taxa é controlada por outro fator, proporcionando a concorrência nas redes. Assim, os nós com maior aptidão tendem a 'vencer' e tornar-se mais bem conectados. Isso explica o porquê de sítios da web, que são atualmente grandes *hubs*, terem crescido tão rapidamente, em poucos anos, relativamente aos demais.

2.5.3 Análise de redes complexas

Muitas ideias atualmente usadas na análise de redes foram inicialmente introduzidas das Ciências Sociais e advindas de uma disciplina chamada Análise de Redes Sociais (ARS), conhecida na literatura internacional por *Social Network Analysis* (SNA). Os métodos e técnicas da ARS agora estão em amplo uso na análise de redes nas mais diversas áreas do conhecimento (NEWMAN, 2010). Observa-se que mesmo para a análise de redes complexas, que investiga redes mais abrangentes do que da análise de redes sociais, é vantajoso estudar os métodos da ARS com grandes possibilidades de generalizar seus resultados para a análise de redes em outras áreas. Há muita literatura disponível sobre a ARS, que estuda as relações sociais entre um conjunto de atores (MIKA, 2007), com características fortemente interdisciplinares (WASSERMAN; FAUST, 1994). A ARS auxilia a detecção e interpretação de padrões nas ligações entre os atores da rede (NOOY; MRVAR; BATAGELJ, 2011), e pode ser empregada de forma bastante exploratória.

Os fenômenos que ocorrem em uma rede não dependem predominantemente das características intrínsecas – chamadas de atributos – de seus nós (FRANCO, 2012), mas de toda a formação topológica da rede. Do mesmo modo, uma rede não se reduz a uma simples soma de suas relações, pois sua forma topológica exerce uma influência sobre essas relações (DEGENNE; FORSE, 1994 apud MARTELETO, 2001). De fato, segundo Kadushin (2004), as teorias fundamentadoras da ARS é uma das poucas, senão a única teoria da ciências sociais que não é reducionista. Ou seja, as redes devem ser analisadas como um todo, pois as propriedades de suas partes não necessariamente explicam o todo. De modo geral, essa

afirmação pode ser generalizada para outros tipos de redes, tendo em vista que a topologia das redes é determinante em muitas situações de análises.

A análise de redes sociais em sua modalidade denominada de exploratória é feita em quatro etapas (NOOY; MRVAR; BATAGELJ, 2011):

(i) a definição das redes;

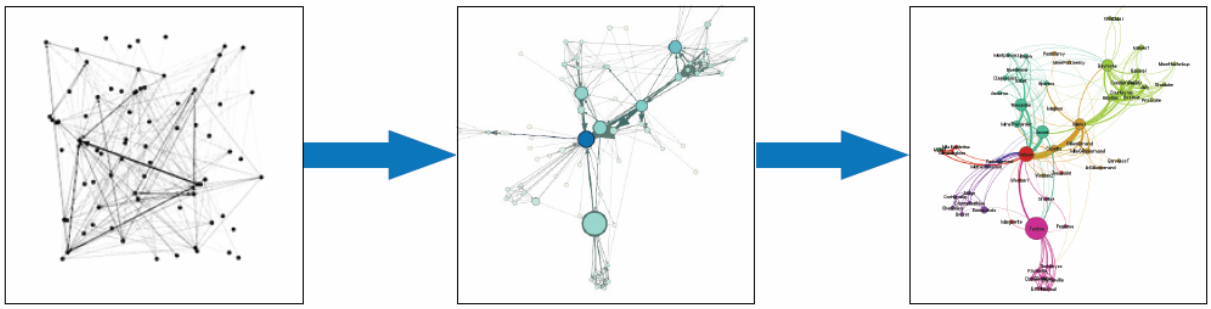
(ii) a manipulação dessas redes;

(iii) a determinação de características estruturais das redes; e

(iv) sua inspeção visual.

A inspeção visual das redes sociais analisadas de forma exploratória é muito útil, por permitir a visualização imediata de características estruturais importantes da rede, ao invés de simplesmente analisar os seus dados, sem uma visualização espacial. A visualização exploratória de redes sociais é facilitada pela disponibilidade de vários algoritmos para geração automática de *layouts* de redes. O olho humano é extremamente talentoso em escolher padrões, e assim, as visualizações de grafos com layouts automaticamente gerados permitem colocar esta habilidade para trabalhar em problemas de rede (NEWMAN, 2010). O uso da inspeção visual requer antes um tratamento na formatação da rede e o uso de métricas para destacar elementos que antes não eram facilmente visíveis, como mostra o exemplo da Figura 23.

Figura 23 – Etapas de tratamento de uma rede para inspeção visual



Fonte: < https://gephi.org/tutorials/gephi-tutorial-quick_start.pdf>. Acesso em: 10 fev. 2016.

Por outro lado, Newman (2010) alerta que a visualização direta de redes só é realmente útil para redes de até algumas centenas ou milhares de vértices. Além disso, explorar a estrutura de uma rede por cálculo é muito mais conciso e preciso do que uma inspeção visual. No entanto, em alguns casos, a exploração por cálculo pode ser abstrata e de difícil interpretação (NOOY; MRVAR; BATAGELJ, 2011). Além disso, a análise de redes não constitui um fim em si mesma, mas um meio para realizar uma análise estrutural com

objetivo de elucidar fenômenos em que a forma da rede explica os fenômenos analisados (MARTELETO, 2001).

Cientistas de uma grande variedade de campos de estudo, ao longo dos anos, desenvolveram um extenso conjunto de ferramentas de matemática, computação e estatística, para análise, modelagem e entendimento das redes (NEWMAN, 2010). Existe uma boa quantidade de softwares, ferramentas e bibliotecas computacionais para auxiliar na análise de redes, seja fazendo medidas e produzindo métricas ou trabalhando na apresentação visual de redes por intermédio de algoritmos apropriados. Alguns desses softwares são:

(i) Gephi[52]: simples, interativo e bom para visualização dinâmica de redes;
(ii) NetLogo[53]: bom para construir simulações de redes;
(iii) R[54]: robusto e na modalidade de programação por scripts;
(iv) iGraph[55]: para cálculos mais sofisticados e programação;
(v) Pajek[56]: robusto para trabalho com redes grandes;
(vi) UCINet[57]: focado em funcionalidades sociológicas,
(vii) NodeXL[58]: realiza análise de redes integrado em planilha eletrônica;
(viii) NetworkX[59]: é um pacote para a linguagem de programação Python; e
(ix) SoNIA[60]: bom para acompanhar a evolução das redes ao longo do tempo.

Na próxima subseção são apresentadas algumas medidas e métricas para auxiliar na análise de redes complexas.

2.5.4 Medidas e métricas de rede

Através da análise da estrutura de uma rede, pode-se calcular uma variedade de medidas úteis que representam especificidades de sua topologia (NEWMAN, 2010). Uma medida ou métrica de rede permite identificar e quantificar a importância de um nó ou um grupo de nós numa rede. Existe na literatura uma grande quantidade de métricas propostas, bem como ferramentas que realizam automaticamente seus cálculos, tais como as

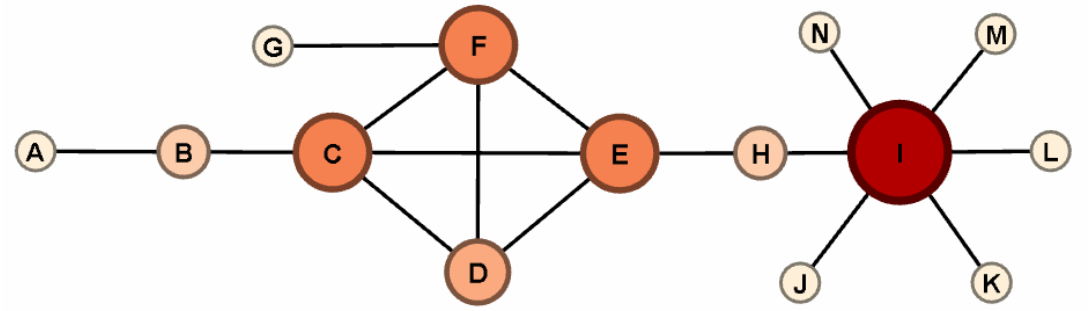
[52] Gephi, disponível em <https://gephi.org/>.
[53] NetLogo, disponível em <https://ccl.northwestern.edu/netlogo/>.
[54] R, disponível em <https://www.r-project.org/>.
[55] iGraph, disponível em <http://igraph.org/r/>.
[56] Pajek, disponível em <http://mrvar.fdv.uni-lj.si/pajek/>.
[57] UCINet, disponível em <https://sites.google.com/site/ucinetsoftware/home>.
[58] NodeXL, disponível em <http://nodexl.codeplex.com/>.
[59] NetworkX, disponível em <https://networkx.github.io/>.
[60] SoNIA, disponível em <https://web.stanford.edu/group/sonia/>.

apresentadas no final da subseção 2.5.3. Nas próximas subseções, são abordadas, de forma breve, apenas as medidas e métricas empregadas no desenvolvimento do presente trabalho.

2.5.4.1 Centralidade de grau (degree centrality) e hub

Na centralidade de grau, ou *degree centrality*, a importância de um nó está no número de ligações que ele possui. A rede da Figura 24 destaca o nó ‘I’ como um *hub*, por ele possuir grau 6, e os nós ‘C’, ‘E’ e ‘F’ são medianamente destacados por possuirem grau 4 cada um.

Figura 24 – Centralidade de grau (*degree centrality*) e hub



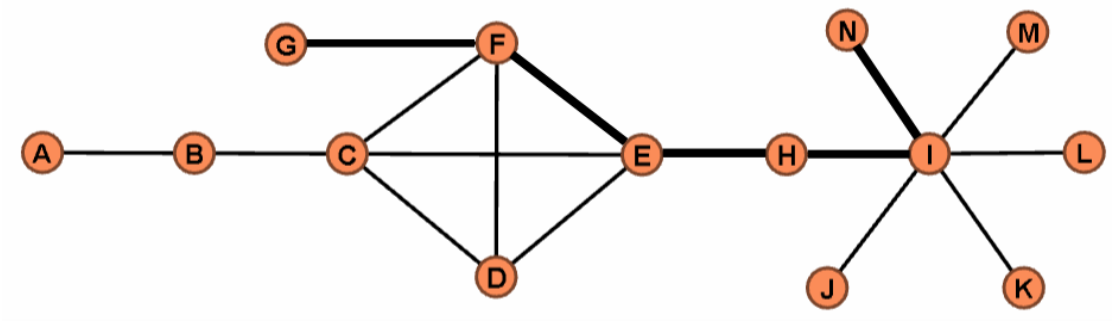
Fonte: Elaboração própria

Em redes direcionadas, as ligações entre os nós da rede tem direção (normalmente representadas por setas). Nesses casos, existem duas variações do cálculo do grau: *in-dregree*, que representa a quantidade de ligações que apontam para o nó e, *out-degree*, que indica a quantidade de ligações que saem do nó.

2.5.4.2 Distâncias

Distância entre dois nós da rede é a menor quantidade de ligações que existem entre esses dois nós. A distância também é chamada de caminho mínimo, caminho geodésico ou *shortest path length*. A Figura 25 representa, de forma destacada, a distância entre o nós ‘G’ e ‘N’. Apesar de existirem outros caminhos maiores, apenas aquele destacado na figura é o mínimo e, portanto, capaz de determinar o valor 5 como distância entre os nós ‘G’ e ‘N’.

Figura 25 – Destaque para distância ou caminho mínimo entre os nós ‘G’ e ‘N’



Fonte: Elaboração própria

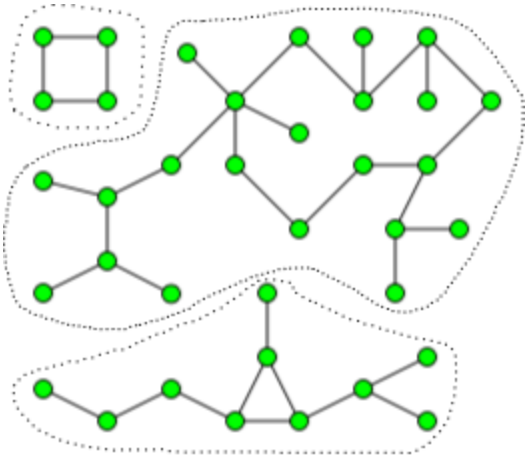
O diâmetro (*diameter)*, de uma rede é o maior caminho mínimo encontrado entre dois nós presentes nessa rede ou, simplesmente, o caminho mínimo entre os dois nós mais distantes. A Figura 25 apresenta uma rede com diâmetro 7, que corresponde à menor distância entre o nó ‘A’ e qualquer um dos nós do conjunto { ‘J’, ‘K’, ‘L’, ‘M’, ‘N’ }.

A média das distâncias, conhecida por *average path length*, é a média de todos os caminhos mínimos possíveis entre todos os pares de nós distintos da rede. Essa medida é útil para saber o quão próximo ou distante os nós estão um dos outros. A rede representada pela Figura 25 tem média das distâncias igual a 3,24 calculada a partir da média das 182 distâncias geodésicas entre os 14 nós distintos da rede.

2.5.4.3 Componente conectado ou connected componente

Componente conectado, ou *connected component*, é um subconjunto de nós de uma rede formado por todos os nós que estão ligados entre si por algum caminho, não necessariamente o mínimo. Na rede da Figura 26 existem exatamente três componentes conectados.

Figura 26 – Rede com três componentes conectados (*connected components*)



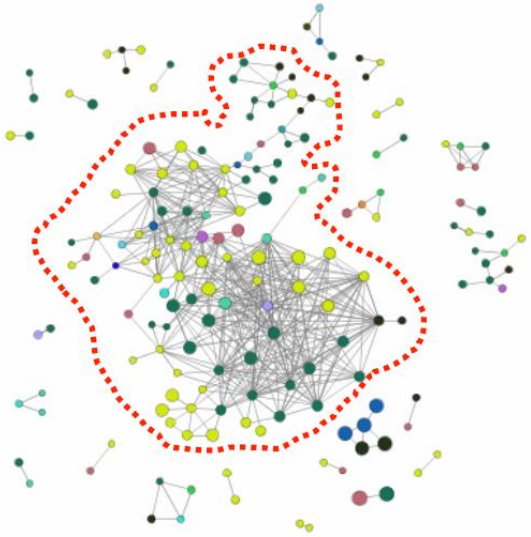
Fonte: adaptado de Wikipedia

2.5.4.4 Componente gigante ou giant component

O componente gigante, ou *giant component*, é o maior componente conectado de uma rede. Porém, normalmente um componente conectado só é considerado componente gigante se o seu tamanho (quantidade de nós) é bem superior ao dos demais componentes conectados porventura existentes na rede. Na rede da

Figura 27 existem vários componentes conectados, contudo, um deles é bem maior e, portanto, considerado componente gigante.

Figura 27 – Rede com componente gigante (*giant component*) em destaque

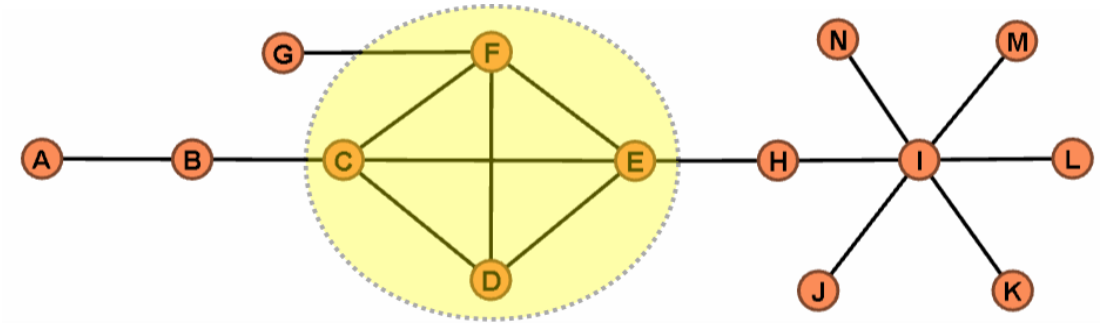


Fonte: Adamic (2013)

2.5.4.5 K-core

Um *k-core* é uma subrede formada pelo subconjunto maximal de todos os nós de uma rede, e suas correspondentes arestas, que apresentam grau superior a *k.* Assim, a Figura 28 destaca os nós ‘C’, ‘D’, ‘E’ e ‘F’, que formam uma rede 3-core a partir da rede original, pois os quatro nós destacados possuem grau maior ou igual a 3.

Figura 28 – Rede na qual se destaca um 3-core



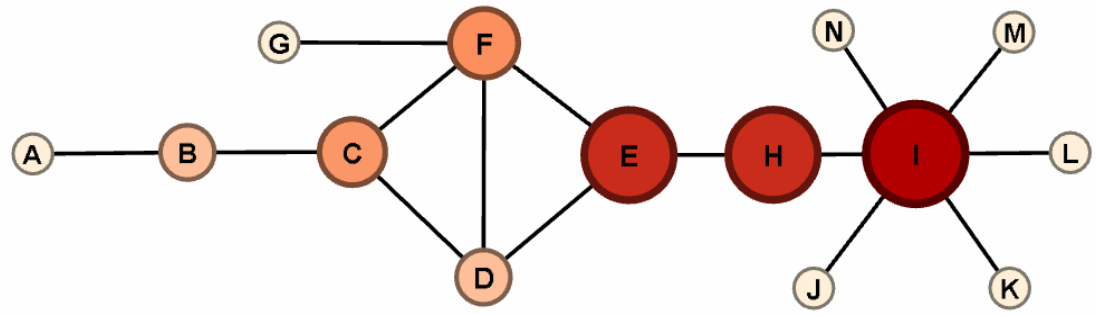
Fonte: Elaboração própria

Observa-se ainda que a execução de um algoritmo para revelar o *k-core* de uma rede exige uma filtragem iterativa e sucessiva de nós com grau inferior a k, até que eles não existam mais. Dessa forma, a subrede destacada na Figura 28, nós ‘C’, ‘D’, ‘E’ e ‘F’, e identificada como 3-core, é também 2-core. Nesse caso houve necessidade de 3 iterações de filtragem de nós com grau inferior a 2, isto é, a primeira filtragem eliminou os nós ‘A’, ‘G’, ‘J’, ‘K’, ‘L’, ‘M’ e ‘N; a segunda filtragem eliminou os nós ‘B’ e ‘I’; a terceira e última filtragem eliminou o nó ‘H’.

2.5.4.6 Centralidade de intermediação ou betweenness centrality

A centralidade de intermediação, ou *betweenness centrality*, mede a importância de um nó quanto à sua capacidade de intermediar o fluxo com os demais nós. A rede da Figura 29 destaca o nó ‘I’ como tendo o maior valor de centralidade de intermediação da rede, e os nós ‘E’ e ‘H’ com valor mediano de centralidade de intermediação.

Figura 29 – Rede com destaque dos nós proporcional à sua centralidade de intermediação (*betweenness centrality*)

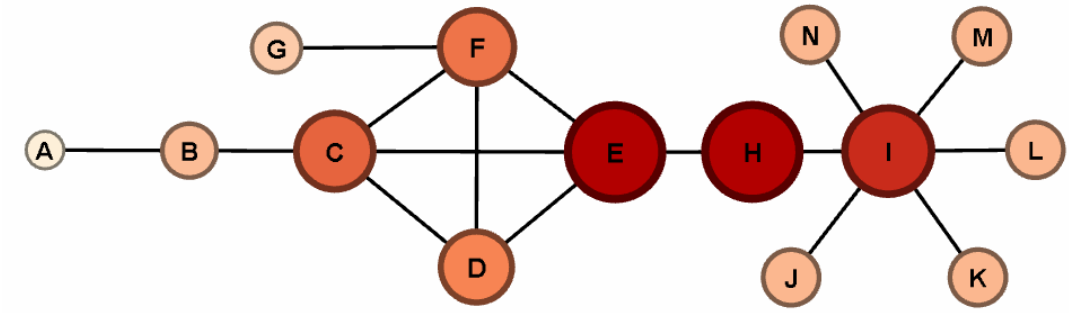


Fonte: Elaboração própria

2.5.4.7 Centralidade de proximidade ou closeness centrality

A centralidade de proximidade, ou *closeness centrality*, mede a importância de um nó quanto à proximidade dele em relação a todos os outros nós da rede. A rede da Figura 30 destaca os nós ‘E’ e ‘H’ como tendo a maior centralidade de proximidade da rede.

Figura 30 – Rede com destaque dos nós proporcional à sua centralidade de proximidade (*closeness centrality*)

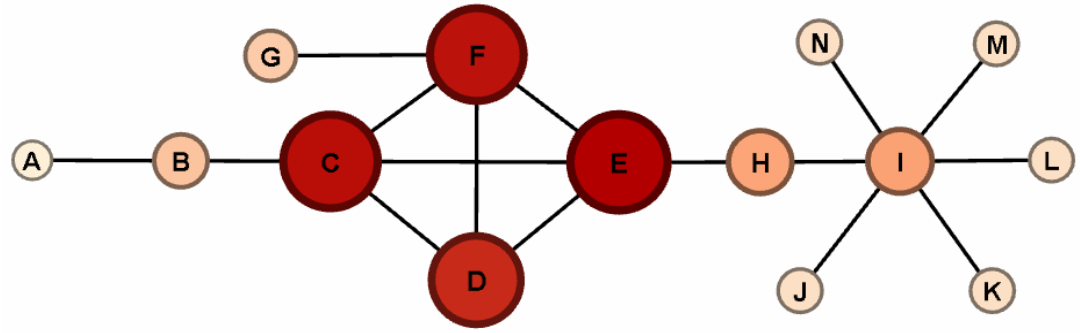


Fonte: Elaboração própria

2.5.4.8 Centralidade de vetor próprio ou eigenvector centrality

A centralidade de vetor próprio, ou *eigenvector centrality*, mede a importância de um nó quanto à quantidade de nós importantes que o referenciam. Assim, a Figura 31 destaca os nós ‘C’, ‘E’ e ‘F’ como os de maiores valores de centralidade de vetor próprio.

Figura 31 – Rede com destaque dos nós proporcional à sua centralidade de vetor próprio (*eigenvector centrality*)

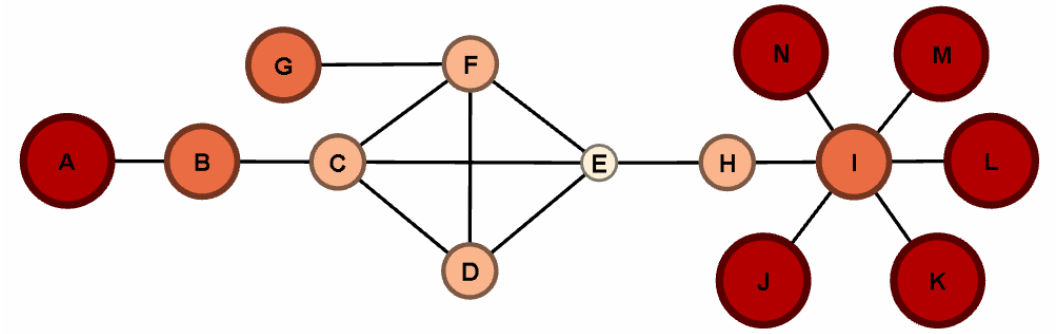


Fonte: Elaboração própria

2.5.4.9 Excentricidade ou eccentricity

A medida de excentricidade, ou *eccentricity*, informa o quanto um determinado nó é excêntrico em relação ao centro de uma rede, ou seja, quanto maior é a distância dele ao centro da rede, maior é a sua excentricidade. A Figura 32 destaca os nós ‘A’, ‘J’, ‘K’, ‘L’, ‘M’ e ‘N’ como os de maiores valores de excentricidade.

Figura 32 – Rede com destaque dos nós proporcional à sua excentricidade (*eccentricity*)

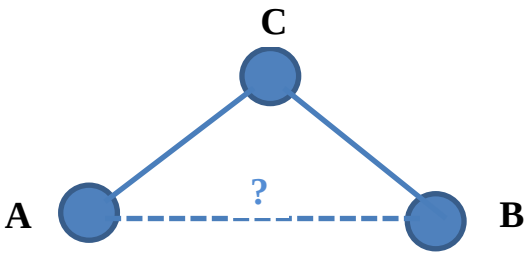


Fonte: Elaboração própria

2.5.4.1 Coeficiente de clusterização

O coeficiente de clusterização mede a probabilidade média de que dois nós vizinhos por outro nó, sejam eles próprios vizinhos diretamente entre si, formando, assim, um triângulo entre os três nós. A Figura 33 representa essa situação, onde o nó A faz vizinhança com o nó B por intermédio do nó C. O coeficiente de clusteriziação verifica, nesse caso, se o nó A possui conexão direta com o nó B.

Figura 33 – Triangulo usado no cálculo do coeficiente de clusterização



Fonte: Elaboração própria

De uma forma geral, essa medida verifica a densidade de triângulos em uma rede. Em redes sociais, essa métrica analisa a típica situação: ‘se os amigos de meus amigos são também meus amigos’.

2.5.5 Considerações finais da seção

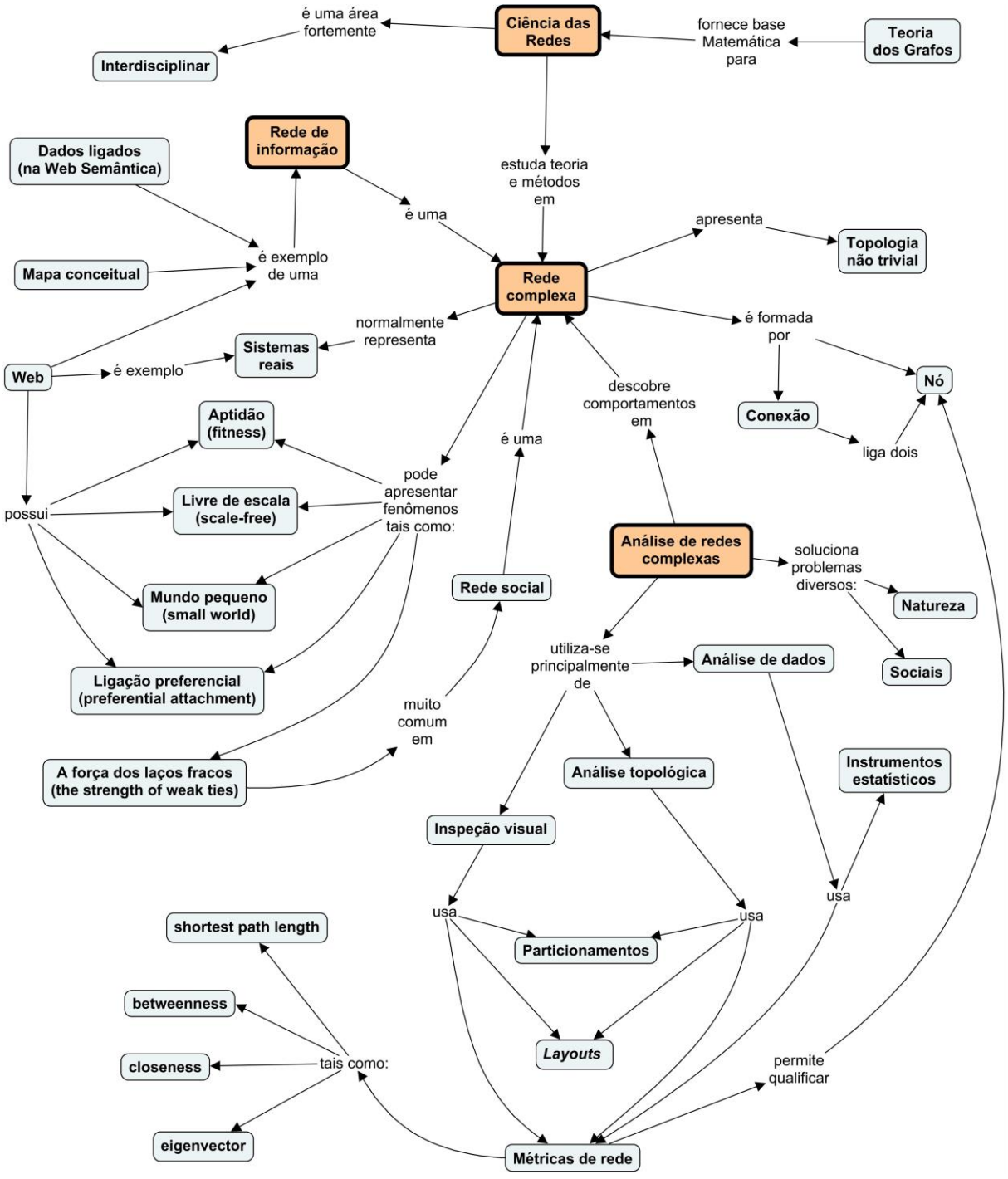
A Ciência das Redes é uma área interdisciplinar, muito ampla e ainda recente. Porém, suas aplicações têm contribuído significativamente para vários campos do conhecimento. Uma de suas vertentes, a análise de redes complexas, permite a descoberta de relações, configurações topológicas, nós com determinadas propriedades em relação ao restante da rede e, sobretudo, a revelação de fenômenos que antes não eram aparentes ou visíveis. Devido a essas revelações, muitos problemas no contexto da natureza e da sociedade podem receber um encaminhamento de solução diferenciado, além de outros problemas e soluções, ainda desconhecidos, que podem emergir.

Existem várias formas de se proceder à analise de uma rede complexa, desde a análise rigorosa de seus dados por meio de instrumentos estatísticos, passando pela averiguação da ocorrência de determinados fenômenos, até a inspeção visual por meio de algoritmos de layout automatizado, aliados ao uso de métricas de rede para fazer ranqueamento de nós como também a sua reorganização por meio de algoritmos de particionamento. Outras formas inovadoras de operar sobre redes estão, atualmente, sendo investigadas por vários pesquisadores ao longo do mundo.

O mapa conceitual da Figura 34 apresenta alguns relacionamentos importantes abordados nessa seção sobre Ciência das Redes, destacando, em cor alaranjada e espessura maior, alguns conceitos relevantes para a presente tese. Entre as várias proposições existentes no mapa, destacam-se aquelas que exemplificam redes de informação, alguns fenômenos que

podem ocorrer em redes complexas, a teoria dos grafos como base para a Ciência das Redes, as aplicações e métodos usados pela análise de redes complexas tais como as métricas de rede.

Figura 34 – Mapa conceitual com alguns relacionamentos abordados na seção 5: Ciência da Redes



Fonte: Elaboração própria

2.6 Considerações finais do referencial teórico

O capítulo do referencial teórico apresentou tópicos e discussões que fundamentam o desenvolvimento da tese. Devido a grande abrangência do trabalho, houve necessidade de discorrer sobre vários assuntos em várias áreas, mesmo que, de forma breve e isolada em alguns casos. Essa seção faz um resgate de relações significativas entre os vários tópicos, buscando formar um corpo coeso de conhecimento, facilitando o seu emprego na metodologia, análise, discussões e conclusões apresentadas nos próximos capítulos.

O termo conceito é importante em vários contextos nesse referencial teórico, principalmente seu uso nos mapas conceituais. As definições propostas por Dahlberg (1978), Teoria do Conceito - subseção 2.2.1, e Novak e Gowin (1984), mapas conceituais - subseção 2.2.3, são compatíveis entre si, pois, sintetizando, descrevem conceito como sendo um enunciado sobre eventos e objetos. Contudo há quem discorde dessa conotação para conceito quando usado no contexto de um mapa conceitual. Dutra, Fagundes e Cañas (2004) defendem a ideia construtivista, de Piaget, onde um conceito é devidamente formado na medida em que ele se relaciona com outros por intermédio de proposições. De qualquer forma, essas duas visões, do conceito relacionado ou do conceito sozinho, não interferem na preocupação unânime da formação de boas proposições no relacionamento entre conceitos no âmbito da construção de um mapa conceitual.

O relacionamento entre conceitos é também muito recorrente nesse trabalho, sendo abordado na subseção 2.2.2.1 que trata sobre a formação de uma proposição, na subseção 2.2.2.2 que aborda os hiperlinks de um hipertexto, na subseção 2.2.3 quando trata das frases de ligação que estabelecem as proposições de um mapa conceitual, na subseção 2.3.3 que discute os dados ligados no contexto da Web Semântica e na subseção 2.5.2.1 que aborda as conexões entre nós de uma rede informacional. As frases de ligação e as proposições são elementos fundamentais nos mapas conceituais. A formação de uma tripla RDF pode ser considerada uma proposição, pois tem dois conceitos, denominados de sujeito/recurso e objeto/valor, que são conectados por uma frase de ligação, denominada de predicado/propriedade.

Nos anos 90 Pierre Levy já indicava sobre a importância dos leitores de um hipertexto conseguirem modificar e acrescentar ligações (LÉVY, 1996). Hoje a web 2.0 concretiza a visão de Levy, pois permite que seus usuários participem e publiquem. Contudo, ela requer aprender a construir em um nova forma não-linear de escrita (CAMPOS; SOUZA; CAMPOS, 2003, p. 8). Os hiperobjetos e hiperinstrumentos, citados na subseção 2.2.2.2, com suas

facilidades em disseminar o conhecimento sem restrições se assemelham ao movimento de dados abertos, da subseção 2.3.2, que também visam o acesso irrestrito a informações, e também os dados abertos ligados, da seção 2.3.3, no contexto da Web Semântica.

Apesar de alguns pesquisadores encararem os mapas conceituais como estruturas hierárquicas (CAÑAS; NOVAK; REISKA, 2015; MOREIRA; MASINI, 1982; SHERRATT; SCHLABACH, 1990), é interessante observar que eles podem ficar mais flexíveis e representar melhor a estrutura cognitiva do seu autor sem essa restrição. Dutra, Fagundes e Cañas (2004) argumentam que são as frases de ligação que estabelecem as fronteiras de um conceito no mapa, segundo a visão Piagetiana de seu trabalho. Dessa forma, a hierarquia deixa de ser importante, uma vez que o autor do mapa pode ainda não ter o sistema hierárquico formado em sua mente e assim não ter elementos mínimos para essa demanda. Lanzing (1997) argumenta que mapa conceitual é uma rede de conceitos. Redes consistem de nós e ligações sem restrição hierárquica e com flexibilidade para conexões que possam ser fieis ao fenômeno ao qual ela representa.

A aproximação da CI com a Web Semântica é inevitável. Um exemplo disso são os movimentos para utilização do padrão RDF da Web Semântica[61] para formatos de dados bibliográficos tal como o MARC21[62]. A Ciência das Redes também se aproxima da Web Semântica quando considera a web e as bases de dados ligados como exemplos de redes de informação, tal como a DBpedia, apresentada na subseção 2.3.5. Chayes (2013) afirmou ser possível modelar a web como um grande grafo finito e, dessa forma, aplicar algoritmos para manipulação de grafos. O conjunto de triplas RDFs que compõe uma base de dados ligados também pode ser considerado uma rede de informação, que por sua vez se equipara a um mapa conceitual. Além disso, o resultado de uma consulta em dados ligados não é apenas um conjunto de *links* para páginas web, mas sim um conjunto de dados estruturados e que podem ser reutilizados em outra aplicação.

Sobre as diversas diferenças entre o conceito de informação, Capurro e HjOrland (2003) sugerem a seguinte reflexão antes de estabelecer qualquer base: que diferença faz se usarmos uma ou outra teoria ou conceito de informação? Dessa forma, destaca-se aqui o conceito de informação formulado por Brookes, nas subseções 2.1.3 e 2.1.4, que encara a informação como sendo um elemento que provoca transformações nas estruturas cognitivas do sujeito, sendo essas estruturas formadas por conceitos que estão ligados entre si e com as

[61] Exemplos de movimentos que organizam informações para conversão de dados bibliográficos para RDF: http://marc-must-die.info/index.php/MARC_to_RDF_mapping e http://librecat.org/
[62] MARC21 é mantido por Library of Congress em http://www.loc.gov/marc/bibliographic/

relações que o indivíduo possui, isto é, a sua imagem de mundo. Nessa linha, a aprendizagem significativa de Ausubel, subseção 2.2.5, é reafirmada, uma vez que o sujeito tem suas estruturas cognitivas modificadas na medida em que seus subsunçores (conhecimento prévio) estabelecem relações com a nova informação, fazendo uma diferenciação progressiva. Esse processo é descrito pela equação fundamental da CI de Brookes, pois a partir de um conhecimento anômalo do sujeito (dúvida ou falta de conhecimento, por exemplo), há uma busca por uma nova informação ΔI, que, entrando em contato com a estrutura cognitiva do sujeito, cria novas relações para a formação de um novo conhecimento.

No contexto da recuperação de informação, Araújo Junior (2007, p. 70) observa que "[...] não existe estratégia de busca a não ser a partir das necessidades de informação dos usuários (com estado anômalo de conhecimento)." Esse mesmo estado anômalo de conhecimento é fundamental para o entendimento da equação de Brookes, pois é a partir dele que o usuário demanda novas informações, que se forem recebidas enquanto documentos recuperados, modificarão sua estrutura cognitiva em conjunto com os seus subsunçores, de acordo com a aprendizagem significativa de Ausubel. Os mapas conceituais, seção 2.2, são uns dos melhores exemplos de aplicação da teoria da aprendizagem significativa, uma vez que, no momento de sua construção, as proposições que relacionam os conceitos, formam uma rede composta de conhecimento prévio e novas informações.

Os mapas conceituais são agentes para a visualização de informação e conhecimento. Hook e Börner (2005) afirmam que um mapa conceitual, de um determinado domínio de conhecimento, adiciona um componente espacial que não estaria disponível em uma apresentação estritamente linear, como um sumário ou uma lista de tópicos navegáveis em um banco de dados. Meyer (2010) cita exemplos de usos de mapas conceituais na visualização de conhecimento. Dentre os exemplos citados, destaca-se o software CmapTools (CAÑAS *et al.*, 2005) que permite aos usuários, individualmente ou em colaboração, representarem, compartilharem e publicarem conhecimento.

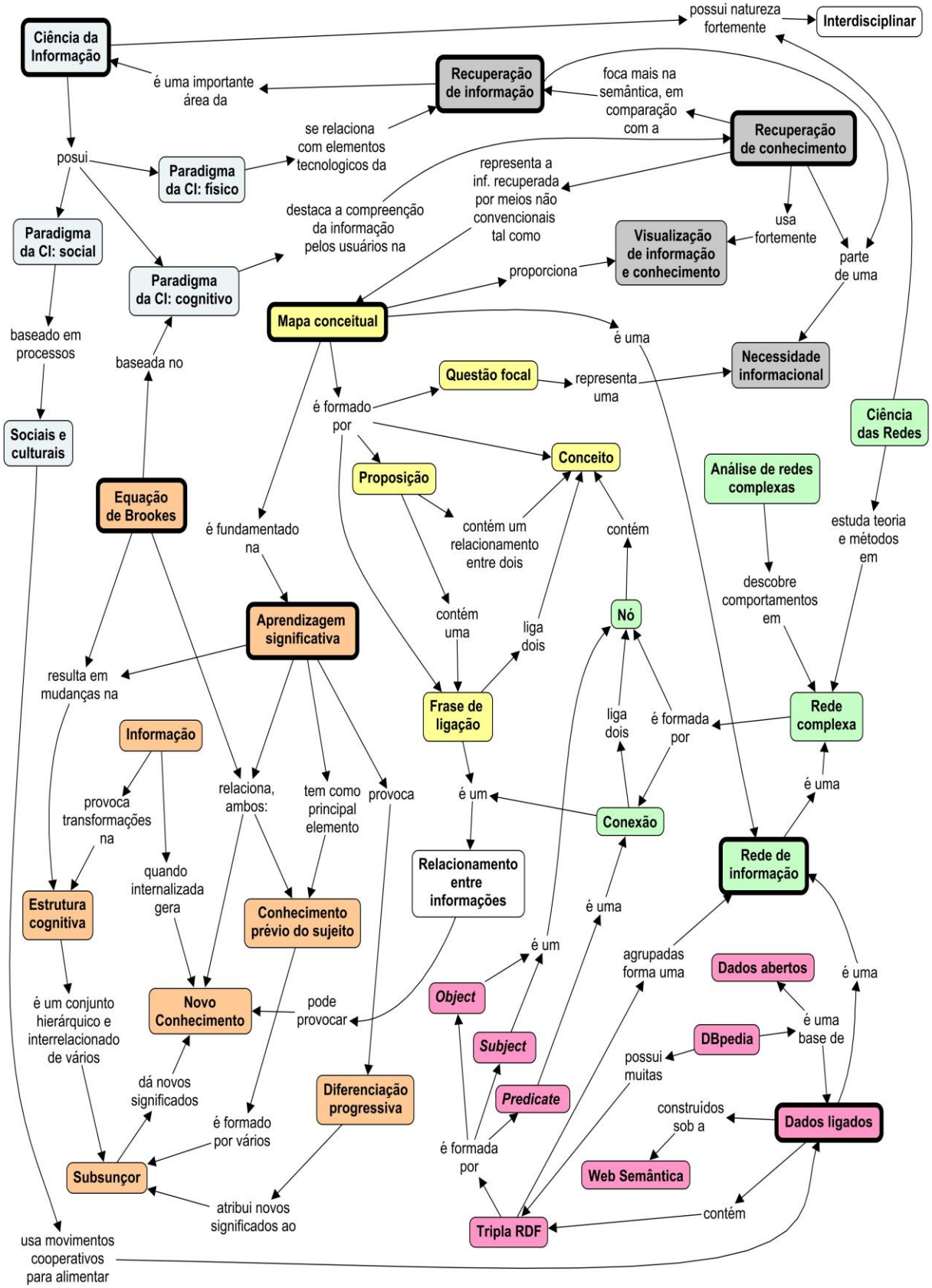
Assim, um mapa conceitual é uma abstração generalista e útil para lidar com o conhecimento de várias formas, fazendo um alinhavo entre vários elementos desse referencial teórico. Ele interliga conceitos por intermédio de relações que constituem proposições bem formadas no estilo ditado por Bertrand Russell. Na equação de Brookes, ele pode ser visto como uma porção de conhecimento capaz de se relacionar com mais facilidade com os subsunçores da estrutura cognitiva do sujeito, isto é, a partir de uma necessidade informacional desse sujeito, registrada no mapa como uma questão focal. O relacionamento desse mapa com a estrutura cognitiva do sujeito, no contexto da aprendizagem significativa,

provoca uma diferenciação progressiva que, na medida em que outras modificações acontecem, gera uma reconciliação integradora. Enquanto rede de informação, o mapa conceitual se aproxima conceitualmente das redes formadas pelas triplas de RDFs das bases de dados ligados. O mapa conceitual é também uma abstração que colabora na visualização de informação e conhecimento e na última etapa no processo de recuperação de conhecimento onde a informação recuperada é apresentada ao usuário. Finalmente, enquanto rede de informação, o mapa conceitual se aproxima conceitualmente das redes formadas pelas triplas de RDFs das bases de dados ligados.

A Figura 35 mostra um mapa conceitual com a questão focal ‘Como se relacionam os assuntos do referencial teórico?’. Considerando a ampla quantidade de assuntos abordados no referencial teórico dessa tese, foram escolhidos alguns relacionamentos importantes e que, de certa forma, sintetizam os mapas conceituais já apresentados no final de cada uma das cinco seções, por intermédio da Figura 2, Figura 4, Figura 10, Figura 20 e Figura 34. Alguns conceitos fundamentais nesses relacionamentos foram destacados com uma espessura maior na caixa: ‘Ciência da Informação’, ‘Recuperação de informação’, ‘Recuperação de conhecimento’, ‘Mapa conceitual’, ‘Rede de informação’, ‘Equação de Brookes’, ‘Aprendizagem significativa’ e ‘Dados ligados’. Cada um desses conceitos destacados pertence a um grupo de conceitos semanticamente coesos e identificados com a mesma cor. Sendo que os conceitos ‘Interdisciplinar’ e ‘Relacionamento entre informações’ estão sem cor, pois ficam na fronteira de grupos. As várias ligações existentes entre esses grupos formam relacionamentos fundamentais que servirão de referência para o restante do trabalho.

Entre as várias proposições existentes nesse mapa relativamente denso pela grande quantidade de conceitos, destacam-se algumas, por exemplo, o forte relacionamento da Ciência da Informação com a área de RI, com a equação de Brookes e com os dados ligados; o ‘conceito’ enquanto elemento intermediador para os subgrupos: mapa conceitual, rede de informação e dados ligados; a relação forte entre a equação de Brookes e a aprendizagem significativa; o mapa conceitual como elemento importante na recuperação de conhecimento e fundamentado na aprendizagem significativa; o conceito ‘relacionamento entre informações’ que representa a frase de ligação de um mapa conceitual, a conexão numa rede de informação tal como os dados ligados, pode provocar novo conhecimento que, se relacionado com o conhecimento prévio do sujeito, provocará mudanças em sua estrutura cognitiva, tal como preconiza a equação de Brookes.

Figura 35 – Mapa conceitual com a questão focal: ‘Como se relacionam os assuntos do referencial teórico?’



Fonte: Elaboração própria

Esse referencial teórico foi desenvolvido para atender aos objetivos da presente tese, elencados na subseção 1.2 da introdução:

- O objetivo geral (desenvolver um modelo para recuperação de informação e conhecimento no contexto dos dados ligados que revele relacionamentos entre termos de uma consulta associada à necessidade informacional do usuário, usando operações de manipulação de redes complexas e geração de mapas conceituais) se relaciona com todos os assuntos desse referencial, com destaque e mais diretamente com a recuperação de informação e conhecimento, seção 2.4, Ciência das Redes, seção 2.5, e mapas conceituais, seção 2.2.
- O primeiro objetivo específico (integrar conceitualmente os temas: Recuperação de Informação e Conhecimento, Redes Complexas, Mapas Conceituais, e Dados Abertos Ligados), por ser uma ação integradora, se relaciona com todos os assuntos desse referencial, principalmente com a seção 2.1 sobre Ciência da Informação que é o elemento coordenador dessa integração;
- O segundo objetivo específico (investigar transformações baseadas na análise de redes complexas que podem ser empregadas na seleção e ranqueamento de relacionamentos em redes de informação, priorizando a descoberta de relacionamentos que satisfaçam uma necessidade informacional) se relaciona mais diretamente com a Ciência das Redes, seção 2.5;
- O terceiro objetivo específico (elaborar um modelo para mapear o fluxo informacional entre dados ligados, redes de informação e mapas conceituais) se relaciona principalmente com os mapas conceituais, seção 2.2, com os dados ligados na Web Semântica, seção 2.3, e Ciência das Redes, seção 2.5;
- O quarto e último objetivo específico (desenvolver e validar junto a um grupo de usuários um protótipo executável computacional que represente o modelo desenvolvido) se relaciona diretamente com a recuperação de informação e conhecimento, seção 2.4, mas recebe contribuições fortes das outras áreas desse referencial teórico: Ciência das Redes, Web Semântica, mapas conceituais e Ciência da Informação.

3 METODOLOGIA

A metodologia da pesquisa é apresentada de forma organizada em cinco seções. A primeira seção caracteriza os métodos empregados de acordo com a literatura da área. As quatro seções seguintes descrevem os métodos empregados nas etapas do trabalho: levantamento bibliográfico e seleção dos trabalhos correlatos; desenvolvimento de um experimento completo de RI para verificar a viabilidade do trabalho, explorar operações de rede e conceber o primeiro modelo de RI; desenvolvimento de um protótipo e o aprimoramento do modelo inicial; e a validação do modelo. Os resultados dessas etapas são descritos no próximo capítulo e a discussão desses resultados, à luz do referencial teórico, é apresentada no capítulo 5.

3.1 Caracterização da pesquisa

O presente trabalho segue o método de pesquisa exploratória que, segundo Gil (2002) e Braga (2007), visa proporcionar maior familiaridade, reunir dados, informações e padrões sobre o problema proposto e através da investigação de relações existentes entre os vários conceitos e processos inerentes ao contexto. A investigação e experimentação de operações de análise de redes complexas para a seleção e ranqueamento de informações em bases de dados ligados foi parte fundamental do processo exploratório dessa pesquisa.

A pesquisa também se caracteriza como qualitativa pelo fato de acontecer uma elaboração de um produto e o seu teste com um grupo de pessoas cujos dados servem para avaliar a adequação aos objetivos propostos (RICHARDSON, 2012) e existir a dependência de interpretações, por parte do pesquisador, do significado desses dados coletados (CRESWELL, 2009). Foi também usado um pequeno aporte quantitativo com o intuito de contribuir para a verificação de informações e facilitar a reinterpretação de observações qualitativas (RICHARDSON, 2012), pois, segundo Pereira (2004), “os dados qualitativos têm uma natureza métrica que, se adequadamente reconhecida por procedimentos de codificação e transformação, permitem seu processamento e análise de modo muito produtivo para a geração de conhecimento” (p. 153). Além disso, quantitativo e qualitativo não se opõem, mas “[...] se complementam, pois a realidade abrangida por eles interage dinamicamente, excluindo qualquer dicotomia” (MINAYO, 2001, p. 22).

Análise de redes é um importante componente do processo metodológico em trabalhos na CI (MATHEUS; SILVA, 2006; SOUSA, 2007; SOUZA; QUANDT, 2008). Amplamente

usada na presente tese, principalmente a parte da inspeção visual por permitir a descoberta imediata de características estruturais importantes, comparar e ranquear métricas das redes de informação construídas a partir das triplas RDFs recuperadas. De posse desses resultados foi concebido um modelo de RI, apresentado na subseção 4.5.

3.2 Primeira etapa: levantamento bibliográfico e trabalhos correlatos

Com base no problema e objetivos, foi feito levantamento bibliográfico sobre os assuntos da tese para compor o referencial teórico. Em seguida investigou-se trabalhos correlatos com a presente tese. A descrição resumida dos trabalhos correlatos bem como o método usado na investigação encontram-se na subseção 1.4 da introdução. Os critérios para a seleção dos trabalhos investigados, bem como a análise comparativa dos trabalhos selecionados com o presente trabalho são apresentados no capítulo de discussão, na subseção 5.8. Esse levantamento bibliográfico e a investigação dos trabalhos correlatos auxiliaram o cumprimento do primeiro objetivo específico elencado na introdução, na subseção 1.2.2.

3.3 Segunda etapa: desenvolvimento do experimento inicial com ciclo completo e concepção da primeira versão do modelo

Com o objetivo de obter dados para verificação da factibilidade inicial da proposta dessa pesquisa foi realizado um experimento com ciclo completo, desde o recebimento de um conjunto arbitrário de termos de consulta do usuário, passando pela geração de redes de informação intermediárias, até a síntese do mapa conceitual resultante. Usando o método de inspeção visual, descrito na seção 2.5.3 do referencial teórico, e atuando de forma experimental no refinamento dos parâmetros de cálculo e transformação executados por cada um dos módulos de software independentes, foi possível obter dados para a concepção da primeira versão do modelo de RI. Os resultados desse experimento bem como a primeira versão do modelo encontram-se na subseção 4.4.

Nessa etapa, várias métricas de redes complexas foram empregadas empiricamente no experimento enquanto método exploratório com o objetivo de descobrir melhores caminhos. As que foram efetivamente adotadas para obtenção dos resultados estão descritas na subseção 4.4. Não houve implementação de software nessa etapa. Foram usados módulos de software independentes e prontos para a execução completa do experimento:

- SPARQL: linguagem de consulta usada para acessar triplas RDFs da DBpedia;
- SNORQL: terminal de consulta usado para acessar dados da DBpedia;
- Gephi: software usado na análise e inspeção visual das redes de informação juntamente com os algoritmos de métricas, particionamento e *layout*;
- Semantic Web Import[63]: software usado para transformar as consultas feitas à DBpedia em redes de informação;
- CmapTools: software usado para formatar e apresentar o mapa conceitual resultante.

Essa etapa auxiliou o cumprimento do segundo e terceiro objetivos específicos elencados na introdução, subseção 1.2.2.

3.4 Terceira etapa: prototipagem e concepção do modelo aprimorado

À luz da equação fundamental da CI de Brookes, apresentada no referencial teórico – subseção 2.1.4, e discutida na subseção 5.1, e baseado nos paradigmas da CI, apresentados no referencial teórico – subseções 2.1.2 e 2.4.1, e discutidos na subseção 5.1.3, e também a partir dos resultados obtidos no experimento realizado, foram concebidos um protótipo para automatizar quase por completo o processo de RI e uma versão aprimorada do modelo de RI. Tanto o protótipo quanto o modelo foram sendo aperfeiçoados em um processo cíclico de investigações e experimentações onde o resultado de um contribuía para o aprimoramento do outro. Esse processo exploratório e cíclico, de desenvolvimento do modelo e do protótipo, consumiu mais de um ano de pesquisa até se atingir um ponto de equilíbrio. Esse método de pesquisa foi determinante para a descoberta de parâmetros de cálculos e transformações de análise de redes que melhorassem empiricamente a síntese do mapa conceitual resultante. O modelo aprimorado, o algoritmo do protótipo bem como os resultados de sua execução são apresentados na subseção 4.5.

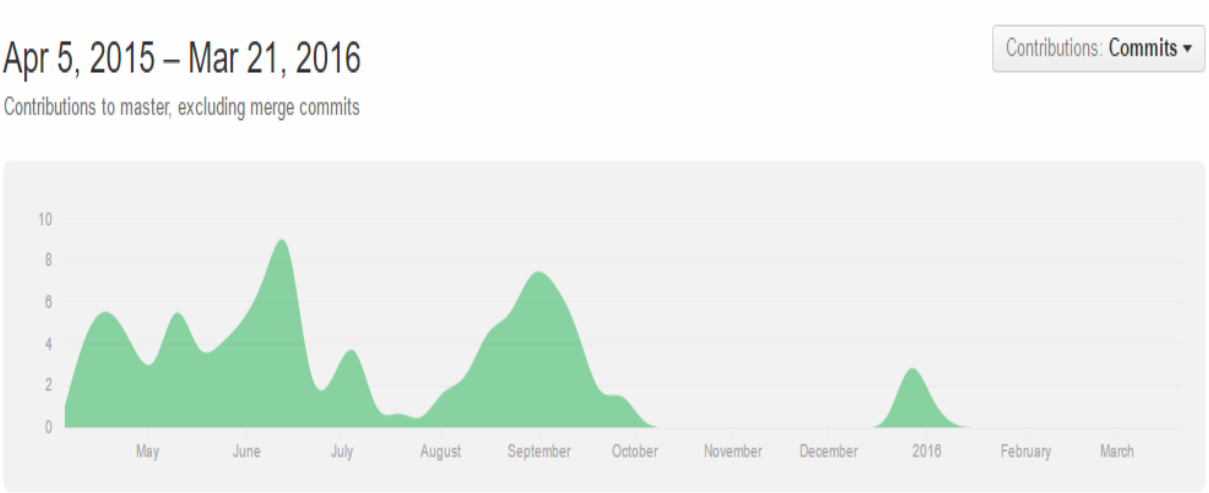
O protótipo permitiu: a execução de maior número de testes; a descoberta de outros elementos de análise de redes complexas; a inclusão de mais iterações no algoritmo que faz a retroalimentação[64] para reiteradas expansões e reduções da rede, até a obtenção do mapa conceitual resultante; flexibilização da quantidade de termos fornecidos pelo usuário; e

[63] Semantic Web Import é um plugin do Gephi que facilita a captura de RDFs em bases LOD e a sua transformação numa rede complexa. Disponível em: <https://marketplace.gephi.org/plugin/semanticwebimport/>.
[64] Retroalimentação é o processo de uso da saída de dados de um sistema na entrada do mesmo sistema.

aumento da alterabilidade, para permitir reconfigurações ágeis do algoritmo. O protótipo também trouxe aprimoramentos adicionais, como a inserção de heurísticas para melhoria da leitura do mapa conceitual resultante, e a possibilidade de validação com um grupo de usuários, apresentado na próxima etapa.

Para a implementação do protótipo foi utilizada a metodologia de Modelo Incremental do Processo, pois, conforme características explicadas por Pressman (2005), existiam requisitos razoavelmente bem definidos e havia necessidade em fornecer um conjunto de funcionalidades antes do término para depois ir refinando e expandindo nas próximas versões. Também foi adotado o Modelo RAD (*Rapid Application Development*), pois, segundo Pressman, a ênfase é num ciclo rápido de desenvolvimento (modelagem: negócio, dados e processo; e construção), possui abordagem de construção baseada em componentes, sendo necessário que os requisitos sejam bem compreendidos e o objetivo do projeto restrito permitindo a criação de um sistema plenamente funcional dentro de um período curto de tempo. Esses dois modelos de desenvolvimento contribuíram para a construção do protótipo, principalmente pela necessidade de testar o protótipo de forma incremental e dinâmica ao longo de todo o processo.

Figura 36 – Frequência de *commits* do desenvolvimento do protótipo



Fonte: Gerado automaticamente pelo sistema GitHub

Foram ao todo sete meses de desenvolvimento em 2015, como pode ser observado na frequência de *commits*[65] da Figura 36 criado automaticamente pelo sistema GitHub[66], que

[65] *Commit*, em Ciência da Computação, é um conjunto de alterações realiza num sistema durante a sua fase de desenvolvimento.

mantém e organiza o repositório no qual o protótipo foi desenvolvido. O código fonte completo do protótipo está organizado em 6 pacotes com 36 arquivos escritos na linguagem Java[67], com 6.157 linhas de código ao todo. Além desses arquivos o projeto ainda possui 5 arquivos fonte entre configurações e modelos. O código completo está disponível para acesso público em https://github.com/hmcristovao/Prototype.

Nessa etapa, foram empregadas métricas de redes complexas que estão descritas na subseção 4.5 bem como, detalhes do algoritmo empregado e exemplos de execuções do protótipo. No desenvolvimento do protótipo foram empregadas as seguintes tecnologias de software:

- Java: linguagem de programação principal usada no desenvolvimento;
- Javacc[68]: usado na geração dos analisadores das entradas de dados, tais como o conjunto de termos, vocabulário controlado, configurações etc.;
- Eclipse[69]: ambiente central e organizador dos vários elementos usados no desenvolvimento;
- Github: usado no controle de versões, armazenamento online e para facilitar possíveis trabalhos cooperativos;
- Egit[70]: ferramenta usada para facilitar o uso do Github dentro do Eclipse;
- Apache Jena[71]: usada na leitura dos dados ligados da DBpedia;
- JSON[72]: formato para auxiliar o intercâmbio de dados entre a base de dados ligados e a biblioteca Apache Jena;
- GraphStream[73]: usada para auxiliar a organização estrutural das redes intermediárias e execução de métricas;
- Gephi Toolkit[74]: usada na execução de métricas e exportação de redes complexas;

---

[66] GitHub é sistema online para hospedagem e organização do trabalho cooperativo de desenvolvimento de software. Disponível em: <https://github.com/>.
[67] Java: linguagem de programa disponível em < http://www.oracle.com/technetwork/java/index.html>.
[68] Javacc é um gerador de compiladores escritos em Java. Disponível em: <https://javacc.java.net/>.
[69] Eclipse é um ambiente integrado de desenvolvimento. Disponível em: <https://eclipse.org/>.
[70] Egit é uma extensão do Eclipse para permitir a integração com o GitHub. Disponível em: <http://www.eclipse.org/egit/>.
[71] Apache Jena é um ambiente para auxiliar desenvolvimentos no contexto da Web Semantica e dados ligados. Disponível em: <https://jena.apache.org/>.
[72] JSON (JavaScript Object Notation) é um formato para intercâmbio de dados, legível por humanos e maquinas. Disponível em: <http://www.json.org/>.
[73] GraphStream é uma biblioteca computacional para manipulação de grafos. Disponível em: <http://graphstream-project.org/>.
[74] Gephi Toolkit é uma biblioteca computacional para manipulação de grafos. Disponível em: <https://gephi.org/toolkit/>.

- CmapTools: usado na geração de layout semiautomático e apresentação visual do mapa resultante,
- Gephi: usado na fase exploratória de investigação e inspeção visual das redes intermediárias.

Essa etapa auxiliou o cumprimento do segundo, terceiro e quarto objetivos específicos elencados na introdução, subseção 1.2.2.

3.5 Quarta etapa: validação do modelo com usuários

Para a etapa da validação do modelo, foi escolhida uma base de dados ligados, selecionado um grupo de usuários, executada a interação do usuário com o protótipo, utilizados instrumentos de coleta de dados e métodos para medição da qualidade da informação recuperada. Essas escolhas e usos são justificados e descritos nas próximas subseções.

Essa etapa auxiliou o cumprimento do quarto objetivo específico elencado na introdução, subseção 1.2.2.

3.5.1 Caracterização da base de conhecimento usada

A base de conhecimento escolhida para suportar os experimentos foi a DBpedia, discutida na subseção 2.3.5 do referencial teórico. Ela é uma base de dados abertos ligados que segue o padrão 5 estrelas de Berners-Lee, isto é, possui os dados sob uma licença aberta, usa URIs para identificar os seus elementos e possui os dados efetivamente ligados entre si e com outras fontes de dados.

Segundo a sua página de estatísticas detalhadas[75], em abril de 2015 existiam 4,5 milhões de entidades padronizadas na língua inglesa contra cerca de 300 mil na língua portuguesa. Em função dessa diferença foi escolhida a língua inglesa para realização dos experimentos com o intuito de proporcionar resultados mais completos.

Fazendo uma análise da DBpedia a luz dos critérios citados por Lancaster (2003) na seção 2.4.5, pode-se destacar a existência de:

- **Cobertura**: considerando uma grande quantidade de triplas RDFs existentes, há uma razoável cobertura sobre assuntos diversos e também ao longo do tempo. Porém,

[75] Página de estatísticas da DBpedia, de abril de 2015: <http://wiki.dbpedia.org/services-resources/datasets/dataset-2015-04/dataset-2015-04-statistics>.

ela está em constante crescimento e é dependente da ajuda coletiva para o seu aumento. Alguns movimentos nesse sentido acontecem na modalidade de *crowdsourcing*.

- **Recuperabilidade**: por intermédio da linguagem SPARQL, a capacidade de recuperação de triplas é completa e bem flexível, porém exige conhecimentos dessa linguagem ou a existência de uma camada de software para tornar mais amigável a comunicação.
- **Previsibilidade**: uma vez recuperada uma tripla, é fácil aferir a importância dela pela sua estrutura simples e lógica contendo dois elementos relacionados por um predicado.
- **Atualidade**: a capacidade em fornecer itens novos ou atualizados é grande uma vez que no momento de sua inserção eles já ficam disponíveis para consultas.

O modelo proposto usa o canal de acesso aos dados RDF da DBpedia, conforme mostrado na Figura 9 (arquitetura da DBpedia) da seção 2.3.5, por intermédio do *SPARQL Clients* com a linguagem de consulta SPARQL e o terminal de consulta (*SPARQL Endpoint*) SNORQL. Isso é, no primeiro modelo proposto, semiautomático onde as consultas foram feitas interativamente pelo terminal SNORQL. No modelo aprimorado, com um grau maior de automatização, o terminal de consulta é substituído pela biblioteca Jena que faz a leitura dos RDFs, e formato JSON para proporcionar o intercâmbio dos dados entre a base de dados ligada e funções da biblioteca Apache Jena. Maiores detalhes na seção 3.4, que trata sobre o protótipo implementado.

3.5.2 Caracterização da amostra de usuários

Tendo em vista a necessidade em avaliar a informação recuperada, os usuários selecionados para participação na validação deveriam satisfazer a dois quesitos importantes:

- Estarem familiarizados com o uso de mapas conceituais. Assim, eles poderiam avaliar a informação recuperada numa ótica independente do entendimento de mapas conceituais enquanto ferramenta para apresentação de conteúdo e, além disso, eles poderiam avaliar o mapa conceitual como resultado da recuperação de informação e ao mesmo tempo como ponto de partida para execução de outras tarefas, tal como, a construção de uma mapa mais completo. Em consonância com esse quesito, Novak e

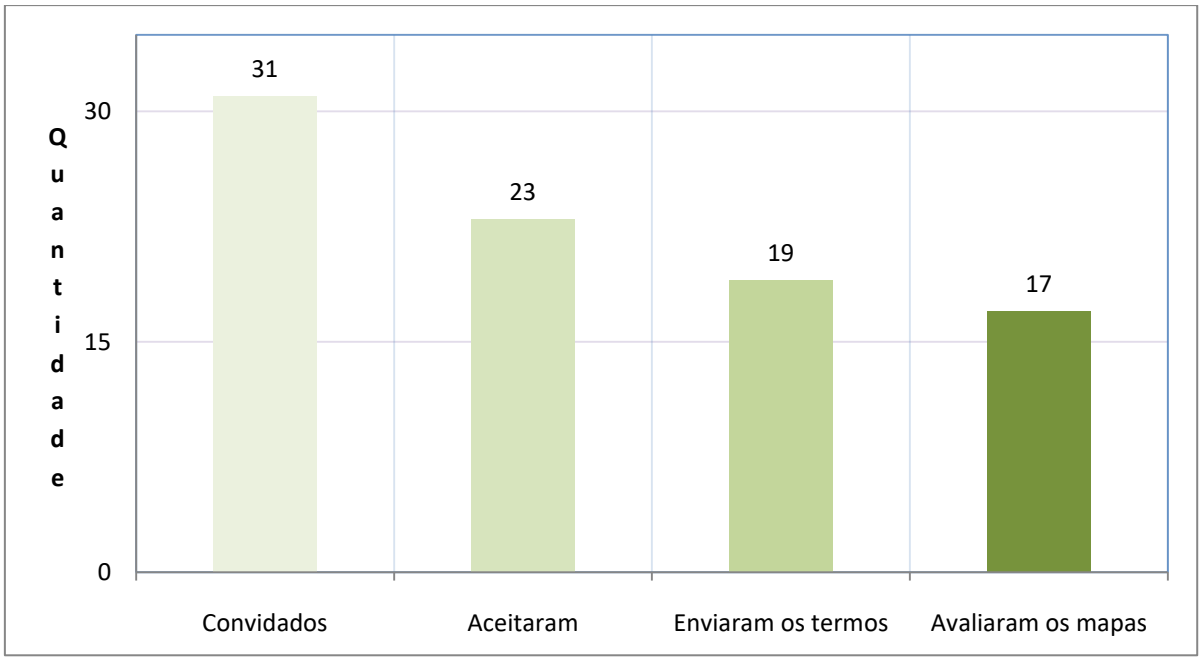
Gowin (1984) alertam sobre a importância da preparação dos aprendizes para lidarem de forma adequada com os mapas conceituais.

- Conhecerem o assunto a ser pesquisado, ou seja, o conjunto de termos fornecido.

Como eles foram solicitados a avaliar a relevância da informação recuperada, esse conhecimento prévio foi importante. Segundo Hjørland (2010), a determinação da relevância de uma informação é fortemente dependente do seu conhecimento. O autor ainda completa dizendo que simples usuários de sistemas de informação não são automaticamente competentes pra julgar relevância da informação tratada pelo sistema.

Em função do cumprimento desses quesitos, foram convidados 31 usuários, entre colegas de trabalho e ex-alunos do autor dessa tese, sendo que todos já trabalharam com mapas conceituais, e a maioria atuou como professor tendo usado mapas conceituais enquanto método de ensino presencial ou à distância. Desses, 23 responderam positivamente e 19 começaram, de fato, a participar da validação enviando os termos de consulta. Porém, apenas 17 terminaram todo o processo, avaliando os mapas conceituais resultantes, como mostra o Gráfico 6.

Gráfico 6 – Quantidade de usuários participantes da validação



Fonte: Elaboração própria

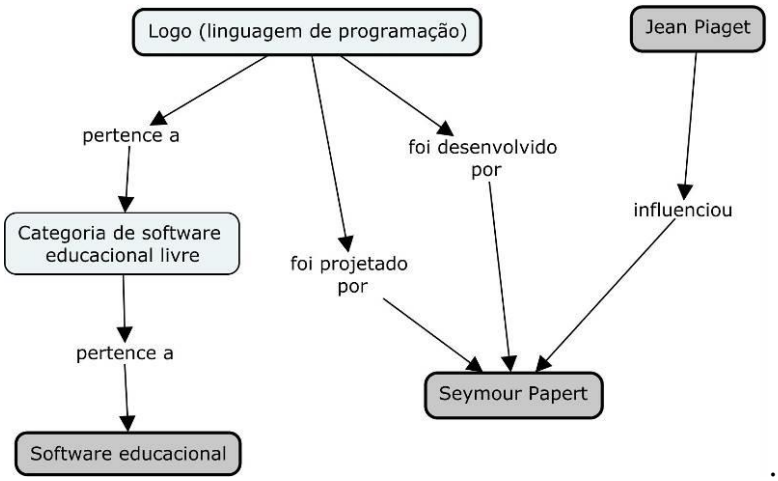
3.5.3 Método de interação com o protótipo

Primeiramente os usuários recebiam mensagem por e-mail explicando o objetivo do trabalho e todo o processo de participação da pesquisa. Em seguida, acontecia o uso do sistema de RI sendo que a interface dos usuários com o protótipo se dava por intermédio de troca de e-mails com o autor dessa tese. Inicialmente os usuários forneciam dois conjuntos de termos semanticamente independentes, com 3 e 6 elementos, e relacionados a uma hipotética necessidade informacional, isto porque uma flexibilização dessa quantidade de termos poderia abrir um espectro muito amplo para análise posterior. Foi também enfatizado aos usuários que o resultado da busca seria em torno da descoberta de relações entre os termos, e não a explicação, detalhamento ou levantamento de características de cada um deles. Os termos poderiam ser fornecidos em português ou inglês, sendo que no primeiro caso eles passariam por uma tradução manual, feita pelo autor dessa tese, com a posterior concordância do usuário. Por conta da base de conhecimento estar em inglês, como já discutido na seção 3.5.1, os usuários receberam orientação desaconselhando o uso de termos difíceis de serem encontrados no contexto da língua inglesa, tais como, nomes de personalidades, localidades e elementos culturais específicos de países que não tivessem o inglês como língua predominante. A relação completa dos termos fornecidos pelos usuários está disponível na seção 4.6.

Após o recebimento dos termos, eles eram processados pelo protótipo até a criação do mapa resultante, sendo que alguns necessitavam de configurações específicas para regulagem do peso das métricas, conforme é discutido na subseção 5.5.2. Como o mapa era criado em inglês, ele passava por uma tradução manual para o português, isto é, se o usuário tivesse enviado os termos em português. Em seguida, os dois mapas conceituais resultantes, referentes aos conjuntos de 3 e 6 termos fornecidos pelo usuário, eram enviados para ele por e-mail. Um exemplo de execução completa, inclusive com as redes intermediárias geradas, encontra-se na subseção 4.5.5. A relação de todos os mapas conceituais resultantes dos termos dos usuários está disponível no APÊNDICE H.

Além desses dois mapas, ele recebia o terceiro mapa conceitual gerado a partir de 3 termos arbitrários, mostrado na Figura 37, que deveria ser avaliado por todos os usuários que conhecessem o assunto abordado pelo mapa. Esse mapa, comum, teve como objetivo a criação de um ponto comum de avaliação e servir de comparação para identificação de usuários com avaliação muito distante dos demais. Além, é claro, de ser mais um indicativo para a validação do modelo.

Figura 37 – Mapa conceitual resultante dos termos ‘Jean Piaget’, ‘Software educacional’ e ‘Seymour Papert’- traduzido



Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

3.5.4 Coleta de dados das avaliações

Ao todo, os 17 usuários realizaram 47 avaliações entre os mapas conceituais gerados a partir dos conjuntos de 3 e 6 termos e o mapa conceitual comum, mostrado na Figura 37. A última fase de participação dos usuários era a avaliação dos mapas conceituais resultantes, por intermédio de um formulário próprio, disponível no APÊNDICE G. Esse formulário, primeiramente, verifica o nível de conhecimento do usuário quanto à leitura e criação de mapas conceituais e o conhecimento sobre o assunto abordado no mapa. Assim, acontece o enquadramento do usuário nos critérios estabelecidos para a amostra, tal como discutido na subseção 3.5.2. Em seguida os usuários avaliaram o quanto o mapa conceitual resultante os auxiliava: (i) no entendimento das relações entre os termos da consulta, (ii) como ponto de partida para uma pesquisa sobre relações com os termos base, (iii) para construir um mapa conceitual mais completo. Foram também avaliadas a relevância dos novos conceitos introduzidos na informação recuperada, intermediários entre os termos enviados na consulta, bem como a relevância das proposições presentes no mapa. Finalmente os usuários avaliaram a completude do mapa, indicando proposições fundamentais que deveriam ter sido recuperadas.

Os dados coletados a partir desse formulário foram tabelados numa planilha eletrônica. Nessa planilha foram computadas as avaliações dos mapas conceituais resultantes e calculados os parâmetros para medição da qualidade da informação recuperada, descritas na

próxima subseção, 3.5.5. Foram também computadas outras informações, tal como a quantidade de triplas RDFs recuperada para cada termo. Os dados coletados são apresentados na seção 4.6 e discutidos nas seções 5.6.1 e 5.6.2.

3.5.5 Método para medição da qualidade baseado nas métricas de RI

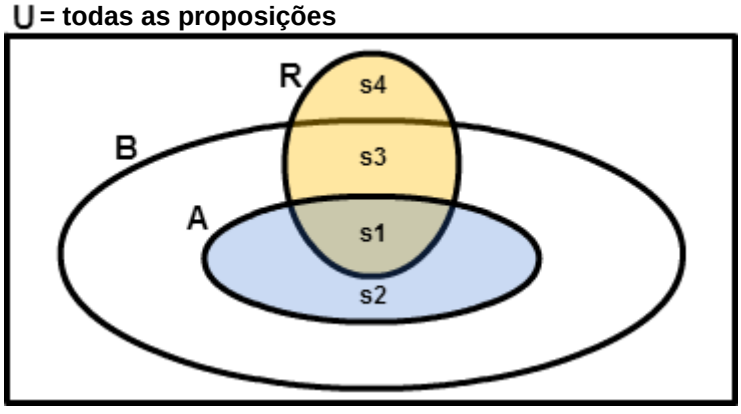
Para medição da qualidade da informação recuperada, foram usadas as medidas ‘precisão’, ‘revocação’ e ‘F-measure’, abordadas no referencial teórico, seção 2.4.5. Como já discutido, elas são medidas bem aceitas para os casos de RI sem interação com o usuário segundo os autores Araújo Junior (2007), Kelly (2009), Baeza-Yates e Ribeiro-Neto (2011) e Kelly e Sugimoto (2013), sendo que Usbeck (2014) - trabalho correlato apresentado na seção 1.4 - também usou essas mesmas medidas para uma recuperação em base de dados ligados.

As seguintes duas subseções apresentam o método de medição usado, sendo que a primeira considera o universo como todas as proposições e a segunda considera o universo como todos os novos conceitos.

3.5.5.1 Contexto das proposições

A Figura 38 é uma adaptação da Figura 15, seção 2.4.5, com o acréscimo do universo formado por todas as proposições e a base de conhecimento. Considera-se que o conhecimento do usuário seja um conjunto que contenha os conjuntos R e A, uma vez que somente serão considerados os questionários onde ele responderá positivamente que conhece o suficiente para poder avaliar.

Figura 38 – Diagrama geral da medição da qualidade da informação recuperada



Fonte: Elaboração própria

Principais conjuntos identificados na Figura 38:

U: universo das proposições

B: base de conhecimento das proposições

R: conjunto das proposições relevantes

A: conjunto de proposições pertencentes à informação recuperada (mapa conceitual resultante)

Proposições da informação recuperada:

s1 U s2 = A

Proposições relevantes e recuperadas:

s1 = R ∩ A

Proposições não relevantes e recuperadas:

s2 = A – R

Proposições relevantes e não recuperadas:

s3 U s4 = R – A

Proposições relevantes, não recuperadas e existentes na base:

s3 = (R – A) ∩ B

Proposições relevantes, não recuperadas e não existentes na base:

s4 = R – B

Proposições relevantes que pertencem à base de conhecimento:

s1 U s3 = R ∩ B

Todas as proposições relevantes:

s1 U s3 U s4 = R

Variáveis provenientes do formulário de avaliação (APÊNDICE G):

Percentual de proposições relevantes no mapa conceitual (questão 4.f).

RelPropPer

Quantidade de proposições relevantes que não apareceram no mapa conceitual (questão 4.g).

RelPropNotShow

Variáveis provenientes de consulta na base de conhecimento:

Quantidade de proposições relevantes que não apareceram no mapa conceitual, e existentes na base de conhecimento.

RelPropNotShowYesBas

Quantidade de proposições relevantes que não apareceram no mapa conceitual, e não existentes na base de conhecimento.

RelPropNotShowNotBas

Variáveis provenientes do mapa conceitual resultante:

Total de proposições no mapa conceitual.

TotProp

Variáveis calculadas:

Percentual de proposições que aparecem no mapa conceitual e que não são relevantes.

NotRelPropPer = 100 - RelProcPer

Total de proposições relevantes no mapa conceitual.

TotRelProp = RelPropPer * TotProp

Total de proposições não relevantes no mapa conceitual.

TotNotRelProp = NotRelPropPer * TotProp

Quantidades:

Quantidade de proposições da informação recuperada:

|s1 U s2| = |A| = TotProp

Quantidade de proposições relevantes e recuperadas:

|s1| = |R ∩ A| = RelPropPer * |A| = RelPropPer * TotProp = TotRelProp

Quantidade de proposições não relevantes e recuperadas:

|s2| = |A – R| = NotRelPropPer * |A| = NotRelPropPer * TotProp = TotNotRelProp

Quantidade de proposições relevantes e não recuperadas:

|s3 U s4| = |R – A| = RelPropNotShow

Quantidade de proposições relevantes, não recuperadas e existentes na base:

|s3| = |(R – A) ∩ B| = RelPropNotShowYesBas

Quantidade de proposições relevantes, não recuperadas e não existentes na base:

|s4| = |R – B| = RelPropNotShowNotBas

Quantidade de proposições relevantes que pertencem à base de conhecimento:

|s1 U s3| = |R ∩ A| + |(R – A) ∩ B| = |R ∩ B |

= RelPropPer * TotProp + RelPropNotShowYesBas = TotRelProp + RelPropNotShowYesBas

Quantidade de todas as proposições relevantes:

|s1 U s3 U s4| = |R ∩ A| + |(R – A) ∩ B| + |R – B| = |R ∩ A| + |R – A| = |R|

= RelPropPer * TotProp + RelPropNotShow = TotRelProp + RelPropNotShow

Métricas para avaliação da qualidade das proposições recuperadas:

Precisão das proposições recuperadas:

Precision = |s1| / |s1 U s2| = |R ∩ A| / |A| = (RelPropPer * TotProp) / TotProp
= TotRelProp / TotProp

Revocação das proposições recuperadas.

Observações:

- Foram consideradas apenas as proposições que estão na base de conhecimento;
- O método para verificar se uma proposição está na base é manual e consiste em abrir os arquivos que representam as redes intermediárias para verificar se há conexão com algum dos conceitos pertencentes à proposição indicada pelo usuário. Se nenhum dos dois conceitos existirem nas redes intermediárias, então executa-se uma consulta SPARQL pelo terminal SNORQL para recuperar as triplas RDFs associadas a um deles para, finalmente, verificar a existência de um relacionamento com o outro conceito.

  RecallTrue = |s1| / |s1 U s3| = |R ∩ A| / |R ∩ B |
  = (RelPropPer * TotProp) / (RelPropPer * TotProp + RelPropNotShowYesBas)
  = TotRelProp / (TotRelProp + RelPropNotShowYesBas)

Média-harmônica das proposições recuperadas:

F-Measure = 2 / ((1/Precision) + (1/Recall)) = 2*(Precision * Recall) / (Precision + Recall)

Estatística do conjunto de usuários para as métricas de avaliação da qualidade das proposições recuperadas:

Média aritmética da precisão das proposições recuperadas:

PrecisionAverage = (somatório das medidas Precision de todos usuários)/qtde de usuários

Média aritmética da revocação das proposições recuperadas (considerando inclusive aquelas que não estão na base de conhecimento) :

RecallAverage = (somatório das med. RecallGeneral de todos usuários)/qtde de usuários

Média aritmética da revocação das proposições recuperadas (considerando apenas as proposições pertencentes a base de conhecimento) :

RecallAverage = (somatório das medidas RecallTrue de todos usuários)/qtde de usuários

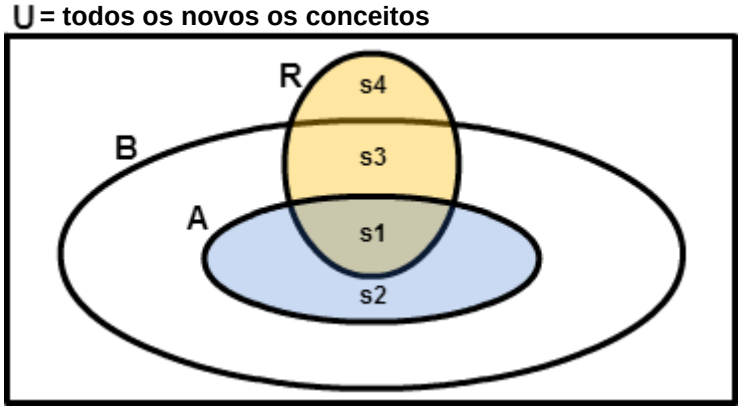
Média aritmética da média-harmônica das proposições recuperadas:

F-Measure Average = (somatório das med. F_measure de todos usuários)/qtde de usuários

3.5.5.2 Contexto dos novos conceitos

A Figura 39 é uma adaptação da Figura 15, seção 2.4.5, com o acréscimo do universo formado por todos os novos conceitos, e a base de conhecimento. Considera-se que o conhecimento do usuário seja um conjunto que contenha os conjuntos R e A, uma vez que somente serão considerados os questionários onde ele responderá positivamente que conhece o suficiente para poder avaliar.

Figura 39 – Diagrama geral da medição da qualidade da informação recuperada



Fonte: Elaboração própria

Principais conjuntos identificados na Figura 39:

U: universo dos novos conceitos

B: base de conhecimento dos novos conceitos

R: conjunto dos novos conceitos relevantes

A: conjunto dos novos conceitos pertencentes à informação recuperada (mapa conceitual resultante)

Novos conceitos da informação recuperada:

s1 U s2 = A

Novos conceitos relevantes e recuperados:

s1 = R ∩ A

Novos conceitos não relevantes e recuperados:

s2 = A – R

Novos conceitos relevantes e não recuperados:

s3 U s4 = R – A

Novos conceitos relevantes, não recuperados e existentes na base:

s3 = (R – A) ∩ B

Novos conceitos relevantes, não recuperados e não existentes na base:

s4 = R – B

Novos conceitos relevantes que pertencem à base de conhecimento:

s1 U s3 = R ∩ B

Todos os novos conceitos relevantes:

s1 U s3 U s4 = R

Variáveis provenientes do formulário de avaliação (APÊNDICE G):

Percentual de novos conceitos relevantes no mapa conceitual (questão 4.e).

RelConcPer

Quantidade de conceitos relevantes que não aparecem no mapa conceitual (questão 4.g).

RelConcNotShow

Variável proveniente do mapa conceitual resultante:

Total de novos conceitos no mapa conceitual.

TotConc

Variáveis calculadas:

Percentual dos novos conceitos que aparecem no mapa conceitual e que não são relevantes.

NotRelConcPer = 100 - RelConcPer

Total de novos conceitos relevantes no mapa conceitual.

TotRelConc = RelConcPer * TotConc

Total de novos conceitos não relevantes no mapa conceitual.

TotNotRelConc = NotRelConcPer * TotConc

Quantidades:

Quantidade de novos conceitos da informação recuperada:

|s1 U s2| = |A| = TotConc

Quantidade de novos conceitos relevantes e recuperados:

|s1| = |R ∩ A| = RelConcPer * |A| = RelConcPer * TotConc = TotRelConc

Quantidade de novos conceitos não relevantes e recuperados:

|s2| = |A – R| = NotRelConcPer * |A| = NotRelConcPer * TotConc = TotNotRelConc

Quantidade de novos conceitos relevantes e não recuperados:

|s3 U s4| = |R – A| = RelConcNotShow

Quantidade de todos os novos conceitos relevantes:

|s1 U s3 U s4| = |R ∩ A| + |(R – A) ∩ B| + |R – B| = |R ∩ A| + |R – A| = |R|
= RelConcPer * TotConc + RelConcNotShow = TotRelConc + RelConcNotShow

Métricas para avaliação da qualidade dos novos conceitos recuperados:

Precisão dos novos conceitos recuperados:

Precision = |s1| / |s1 U s2| = |R ∩ A| / |A| = (RelConcPer * TotConc) / TotConc
= TotRelConc / TotConc

Observação: diferentemente do contexto das proposições, é difícil distinguir conceitos não recuperados que pertencem a base de conhecimento daqueles que não pertencem. Isso porque, no contexto dos dados ligados, um determinado conceito não recuperado poderia estar distante dos conceitos fornecidos pelo usuário. Portanto, não serão calculados a revocação dos novos conceitos recuperados e, por consequência, a média-harmônica dos novos conceitos recuperados.

Estatística do conjunto de usuários para as métricas de avaliação da qualidade dos novos conceitos recuperados:

Média aritmética da precisão dos novos conceitos recuperados:

PrecisionAverage = (somatório das medidas Precision de todos usuários)/qtde de usuários

4 O MODELO HÍBRIDO DE RECUPERAÇÃO DE INFORMAÇÃO E CONHECIMENTO

Esse capítulo apresenta resultados da pesquisa referentes à definição do modelo proposto na sua primeira versão e na versão aprimorada, e também referentes à sua validação. O capítulo é organizado em seis seções: definição do modelo; contextualização do modelo na Web Semântica; descrição dos mapeamentos utilizados; apresentação da primeira versão do modelo a partir do experimento realizado; apresentação do modelo aprimorado juntamente com o algoritmo construído, uma execução completa do protótipo desenvolvido e alguns testes piloto; e os resultados da validação. A discussão dos resultados aqui apresentados encontra-se no próximo capítulo.

4.1 Definição geral do modelo

O modelo de RI proposto no presente trabalho é definido no Quadro 6 baseado na definição de modelo de RI proposta por Baeza-Yates e Ribeiro-Neto (2011), Quadro 4 da seção 2.4.6 do referencial teórico.

Quadro 6 – Quádrupla de definição geral do modelo de RI proposto

O modelo de RI proposto é definido pela quádrupla [**D**, **Q**, ***F***, ***R***(q_i, d_j)] onde:
1. **D** é o conjunto formado pelas triplas RDF de uma base de dados ligados.
2. **Q** é o conjunto formado por todos os grupos de termos que podem ser elencados por um usuário, onde cada grupo representa uma necessidade informacional de um usuário específico. O grupo de termos é também chamado de consulta.
3. ***F*** é o *framework* do modelo que contém: (i) a compilação do formato da consulta de um conjunto de termos para a linguagem SPARQL, (ii) os mapeamentos de dados ligados, triplas RDFs, para o formato de rede de informação e mapa conceitual e (iii) operações de redes complexas, e (iv) processo de retroalimentação dos nós ranqueados para uma nova consulta.
4. ***R***(q_i, d_j) é a função de ranqueamento que associa um conjunto de termos (consulta) $q_i \in$ **Q** e uma tripla RDF $d_j \in$ **D** a uma ordem no ranqueamento geral de todos os RDFs coletados.

Fonte: Elaboração própria

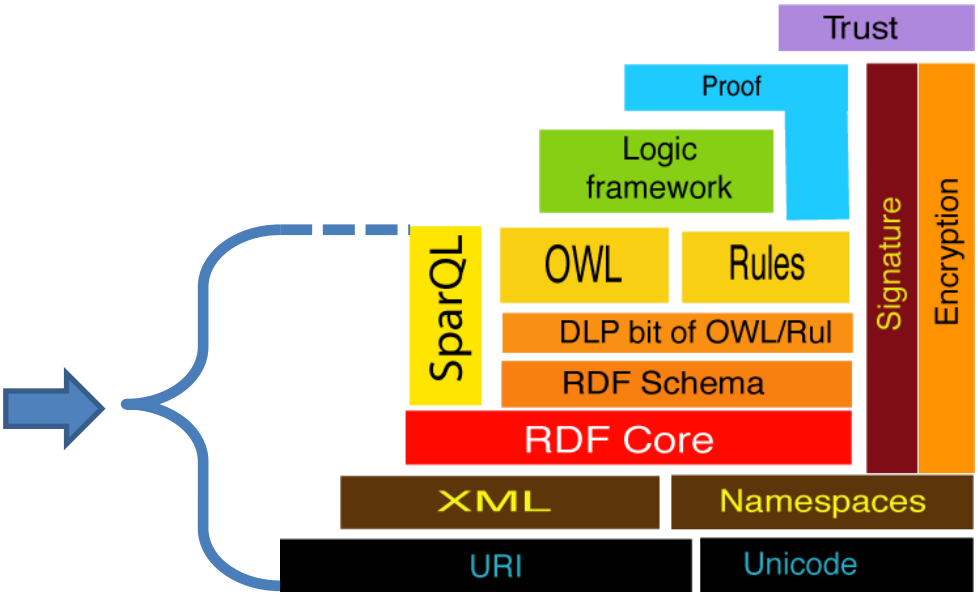
A definição do Quadro 6 especifica a quádrupla [**D**, **Q**, ***F***, ***R***(q_i, d_j)] que representa o modelo de RI proposto. O conjunto **D** é o domínio (*domain*), ou seja, é formado por todas as triplas RDFs da base de dados ligados disponível para o processo de recuperação, como por exemplo, a DBpedia. O conjunto **Q** possui as consultas (*querys*) possíveis de se formular, ou seja, todos os possíveis grupos de termos que o usuários pode determinar para a consulta e que representem o seu desejo informacional. ***F*** é o ambiente (*framework*) do modelo que contém as estruturas de dados usadas, tais como RDFs, redes de informação, tabelas de ranqueamento etc., e os processos usados na seleção e ranqueamento da recuperação, inclusive a função ***R***, quarto elementos da quádrupla. Finalmente, ***R***(q_i, d_j) define a relação (*relation*) que especifica o ranqueamento associando um conjunto de termos escolhido pelo usuário $q_i \in$ **Q** e uma tripla RDF da base de dados ligados $d_j \in$ **D**, a uma posição na ordem do ranqueamento geral que contém todos os RDFs coletados na consulta.

4.2 Interseção do modelo com a Web Semântica

A interceção mais explícita do modelo proposto nessa tese com a Web Semântica é evidenciada no uso dos dados ligados, apresentados no referencial teórico, na subseção 2.3.3. Os dados ligados são usados na formação das bases de conhecimento lidas pelo modelo. Contudo, tendo como base a arquitetura da Web Semântica proposta por Berners-Lee (2005), apresentada na subseção 2.3.1 do referencial teórico, o modelo se enquadra nas camadas destacadas na Figura 40.

A camada URI/UNICODE contém as referências do tipo URI e IRI usadas pelos dados ligados lidos na base de conhecimento. A camada XML/Namespaces permite a comunicação interna no sistema de RI com o uso da linguagem XML que generaliza e potencializa a HTML. A camada RDF Core representa a rede de informações formada pelas triplas RDFs lidas na base de conhecimento. A camada RDF Schema acrescenta taxonomias na camada RDF Core para melhorar o relacionamento entre as informações e o seu acesso. As camadas DLP bit of OWL/Rul e OWL/Rules usam linguagens de ontologia para formalizar, padronizar e melhorar a expressividade semântica das informações e, consequentemente, a interoperabilidade, o seu acesso e uma melhor representação de conhecimento.

Figura 40 – O modelo proposto no contexto da arquitetura da Web Semântica



Fonte: adaptado de Berners-Lee (2005)

4.3 Mapeamentos

Apesar do conjunto de triplas RDFs oriundas da consulta e do mapa conceitual resultante serem considerados redes de informação, foi necessária a realização de mapeamentos para transformações na transição entre cada uma dessas redes de informação. Além disso, adotou-se o uso de um vocabulário controlado para a apresentação do mapa conceitual mais legível.

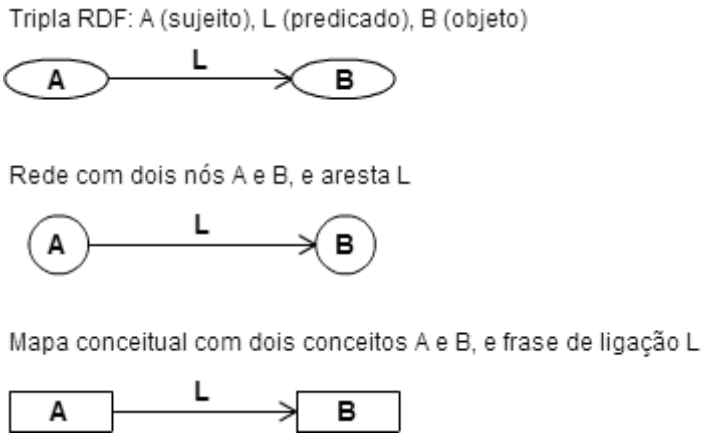
4.3.1 Modelo para mapeamento entre dados ligados, rede de informação e mapa conceitual

O primeiro mapeamento acontece na transição entre a rede formada pelo conjunto de triplas RDFs recuperados e a rede de informação. O segundo mapeamento ocorre na transição entre a rede de informação final e o mapa conceitual resultante. A Figura 41 contém a legenda usada nos dois modelos de mapeamentos, com a distinção gráfica entre as três formas de representação do conhecimento: tripla RDF, nós de uma rede, e a proposição de um mapa conceitual, ou seja, dois conceitos conectados por uma frase de ligação.

A Figura 42 representa o caso geral do modelo para mapeamento, onde o sujeito, objeto e predicado da tripla RDF são transformados, em dois nós conectados em uma rede de informação, posteriormente transformado em dois conceitos com uma frase de ligação em um mapa conceitual. A Figura 43 representa também a transformação de tripla RDF para rede e,

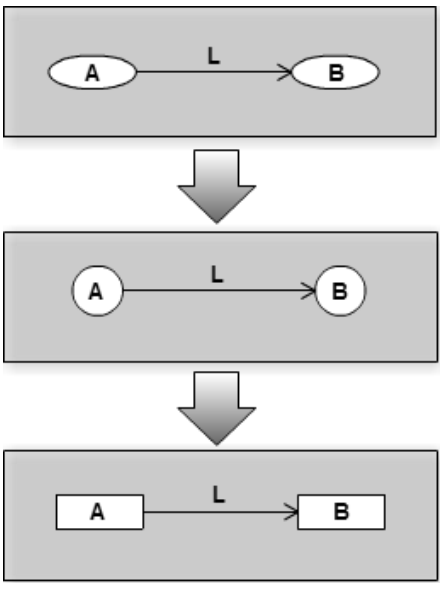
depois, para mapa conceitual, porém considerando a situação de um grupo de triplas RDFs onde os sujeitos são iguais e os predicados e objetos são diferentes. Nesse caso, os sujeitos são unificados em apenas um nó da rede e um conceito do mapa conceitual. Existem ao todo 11 casos no modelo para mapeamento. Os outros 10 casos encontram-se no APÊNDICE A.

Figura 41 – Legenda do modelo para os mapeamentos



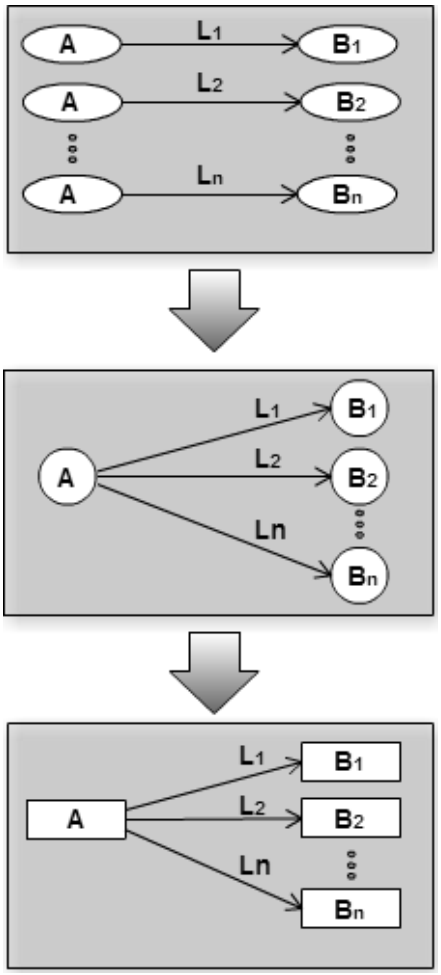
Fonte: Elaboração própria

Figura 42 – Caso geral



Fonte: Elaboração própria

Figura 43 – Sujeitos iguais, predicados e objetos diferentes



Fonte: Elaboração própria

4.3.2 Vocabulário controlado

Tal como discutido na seção 2.3.4 do referencial teórico, o vocabulário controlado é uma estruturação do conhecimento capaz de inibir disparidades, problemas de sinônimos e grafia errada e, sobretudo, um acordo sobre o significado de um conjunto de termos usados em um sistema. Dessa forma, no presente trabalho, ele é uma relação de termos que representam os predicados das triplas RDFs com as suas respectivas transições, usando o modelo de mapeamento da subseção 4.3.1, para as frases de ligação do mapa conceitual resultante com o intuito de aumentar o nível de legibilidade do mapa. Exemplos: o predicado '*schoolTradition*' é transformado na frase de ligação '*has school tradition*', o predicado '*field*' é transformado na frase de ligação '*works in the field*'. O vocabulário parcial usado na execução do protótipo encontra-se no APÊNDICE F.

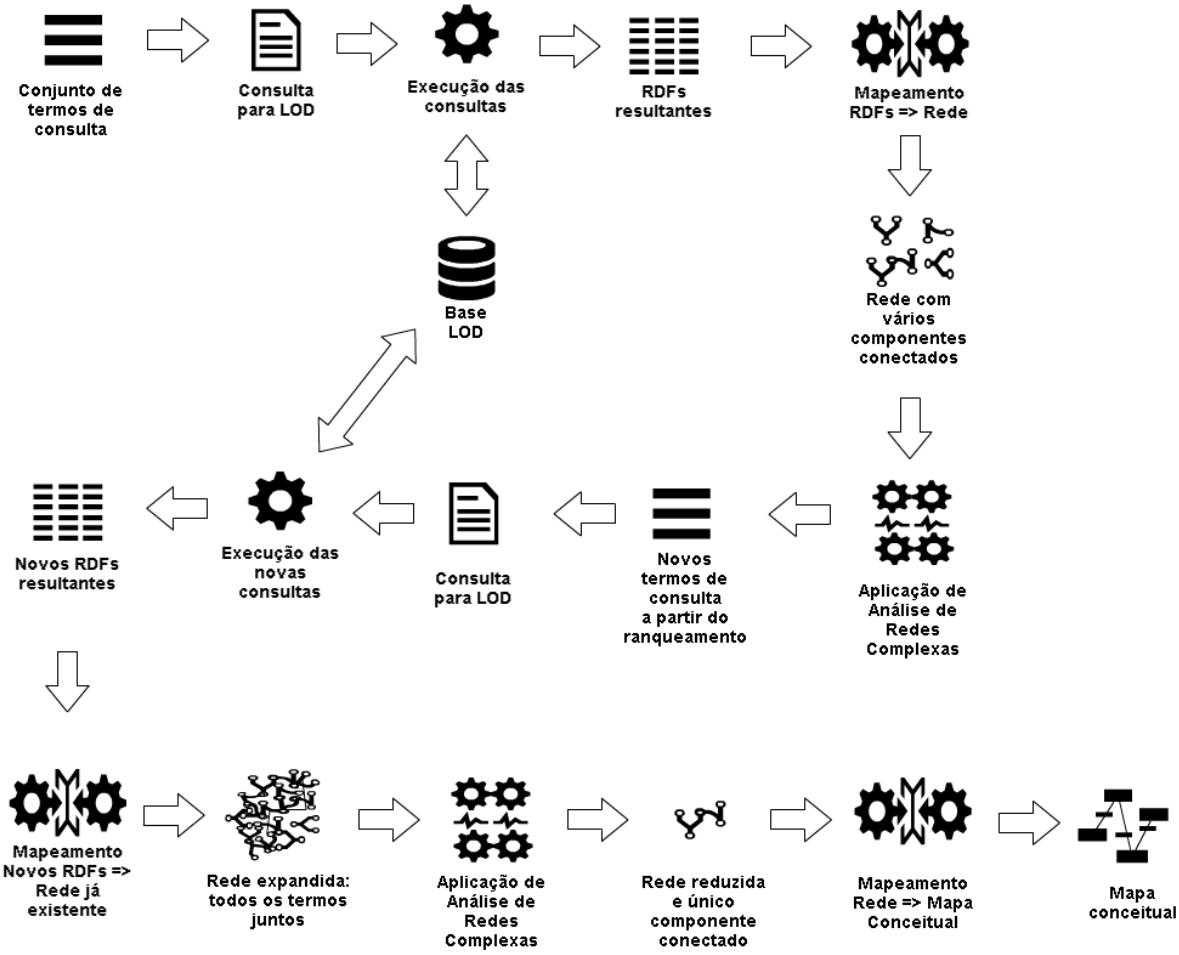
4.4 Experimento inicial com ciclo completo

A realização experimental de um ciclo completo do processo, desde a entrada dos termos de consulta até a geração do mapa conceitual resultante, foi uma etapa importante no desenvolvimento do trabalho como um todo, tal como já discutido na metodologia, seção 3.3. A presente seção mostra os resultados desse experimento, começando pela primeira versão do modelo, o processo completo da exploração realizada nas redes de informação criadas pela recuperação das triplas RDFs na base de conhecimento de dados ligados, a construção do mapa conceitual resultante desse experimento, e as conclusões e problemas revelados no experimento com o intuito de subsidiar a construção do modelo aprimorado descrito na próxima seção.

4.4.1 Primeira versão do modelo

A Figura 44 sintetiza o processo completo da primeira versão do modelo. Observa-se como início do processo, a entrada do conjunto de termos arbitrários a serem buscados por intermédio de uma consulta na base de dados abertos ligados (*linked open data* – LOD) escrita em SPARQL. Em seguida ocorre a execução dessa consulta em um terminal SNORQL sobre a base LOD. O conjunto de RDFs resultantes da recuperação passa por um mapeamento para a formação de uma rede informacional, normalmente com vários componentes conectados. A aplicação da análise de redes complexas é uma das partes mais importantes pois é quem revela novos termos a partir dos ranqueamentos realizados. Uma nova consulta para LOD é montada com a sua execução sobre a base LOD e o retorno de novos RDFs resultantes. Esses novos RDFs são mapeados e mesclados na rede de informação anterior para a formação da rede expandida. Uma nova exploração com a análise de redes complexas acontece, agora, com o intuito de diminuir a rede reduzindo-a a um único componente conectado. Finalmente acontece o mapeamento da rede de informação final no mapa conceitual resultante.

Figura 44 – Diagrama geral do primeiro experimento realizado.



Fonte: Elaboração própria

A construção da parte referente à formação da rede intermediária, análise de redes complexas e a construção do mapa conceitual resultante no modelo é detalhada nas próximas subseções.

4.4.2 Construção da primeira rede de informação

Como ponto de partida do experimento, foram escolhidos oito termos de busca presentes na DBpedia: ‘Information Science’, ‘Education’, ‘Psychology’, ‘Semantic web’, ‘Jean Piaget’, ‘Lev Vygotsky’, ‘David Ausubel’ e ‘Tim Berners-Lee’. Como esse experimento inicial e completo levaria muito tempo para ser desenvolvido, predominantemente por processo manual, fez-se uma escolha cuidadosa de termos que levou em consideração os seguintes critérios: (i) mesma quantidade de pessoas e conceitos; (ii) áreas de conhecimento diversas, porém bem conhecidas; (iii) termos com quantidade variada de

resultados, isto é, quantidade de triplas RDFs resultantes da busca na DBpedia em maior número e menor número; (iv) quantidade inicial de termos de busca com meta para um mapa conceitual resultante com cerca de 20 conceitos, tornando-o de fácil leitura; e (v) alguns termos com relacionamentos bem conhecidos entre si e outros sem uma relação aparente.

Pelo terminal SNORQL foi escrita uma consulta na linguagem SPARQL para extrair da DBpedia as triplas RDFs onde cada um dos 8 termos pudesse fazer o papel de *subject* ou de *object* numa tripla. Assim, foram recuperadas 4.697 triplas RDFs, parcialmente mostradas na Figura 45, onde a coluna ‘s’ representa o *subject* da tripla RDF, a coluna ‘p’ o *predicate* que estabelece o relacionamento entre o *predicate* e o *object*, e a coluna ‘o’ o *object*.

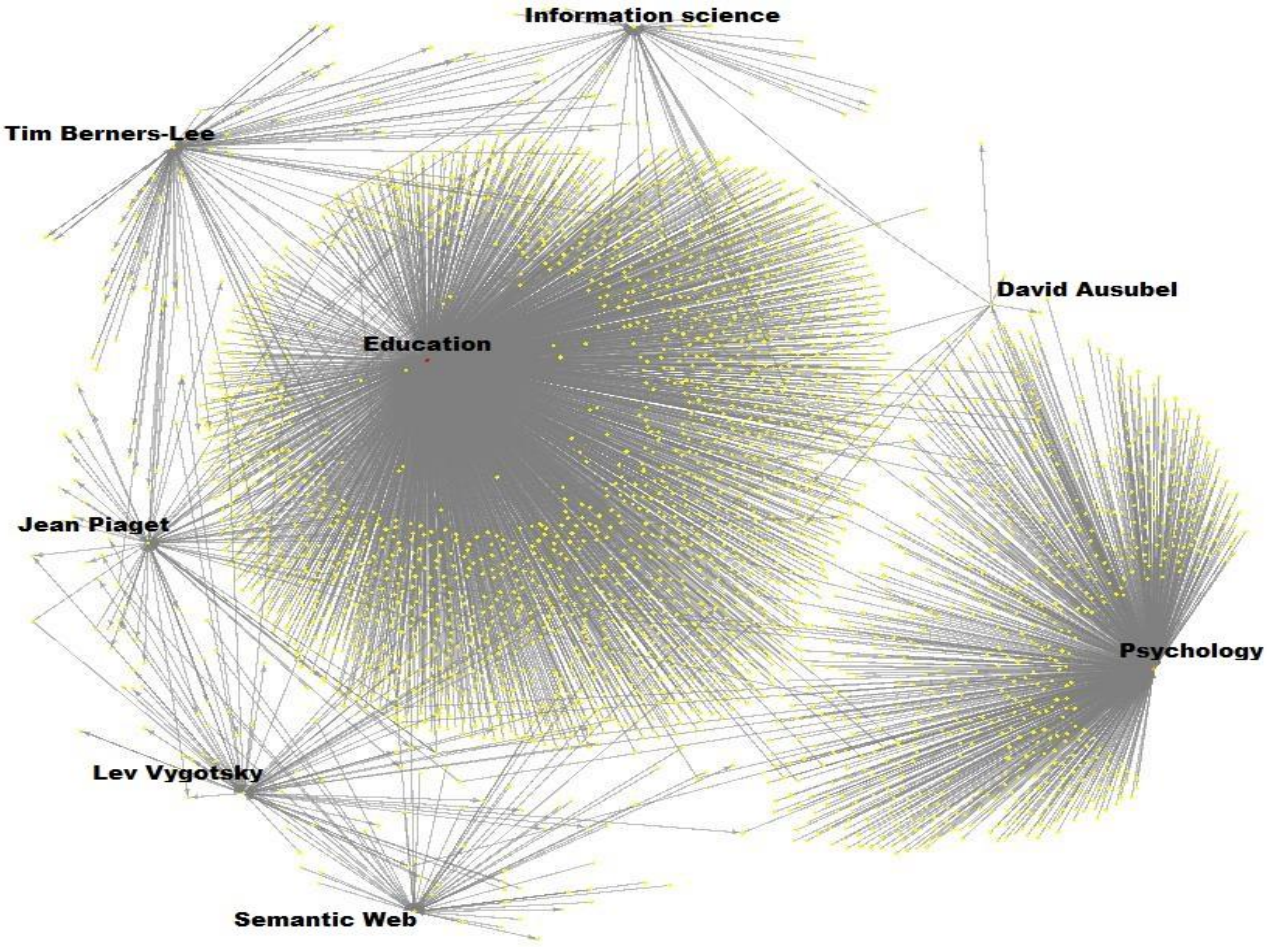
Figura 45 – Alguns dos 4.697 RDFs resultantes da consulta dos 8 termos à DBpedia, pelo terminal SNORQL

s	p	o
:Wellspring_Retreat_and_Resource_Center	dbpedia:ontology/genre	:Psychology
:Benjamin_Rush	dbpedia:ontology/occupation	:Education
:Jacob_L._Moreno	dbpedia:ontology/field	:Education
:Advanced_Technical_Intelligence_Center	dbpedia2:industry	:Education
:College_of_Fine_Arts,_Thrissur	dbpedia:ontology/type	:Education
:Lev_Vygotsky	<http://purl.org/dc/terms/subject>	:Category:Deaths_from_tuberculosis
:Denis_Pelli	dbpedia2:field	:Psychology
:Jean_Piaget	dbpedia:ontology/influencedBy	:Henri_Bergson
:Education	<http://purl.org/dc/terms/subject>	:Category:Philosophy_of_education
:Eugene_Galanter	dbpedia2:field	:Psychology
:Jean_Piaget	<http://purl.org/dc/terms/subject>	:Category:Mathematical_cognition_researchers
:Tim_Berners-Lee	dbpedia:ontology/award	:Royal_Academy_of_Engineering
:Information_science	<http://gephi.org/homepage>	<http://en.wikipedia.org/wiki/Information_science>

Fonte: Elaboração própria

O próximo passo foi o uso do software Gephi com apoio do plugin Semantic Web Import, que fez a transformação dos 4.697 RDFs extraídos da base DBpedia para uma rede direcionada composta de 2.854 nós e 2.900 arestas. A Figura 46 mostra a rede obtida após mapeamento do conjunto de triplas RDFs, onde as triplas que compartilham sujeitos ou objetos com outras triplas constituem a rede. Para a formatação da rede, Figura 46, Foi aplicado o *layout* Force Atlas 2, disponível no software Gephi, com destaque para os oito nós que representam os termos escolhidos.

Figura 46 – Rede de informação obtida após mapeamento do conjunto de RDFs



Fonte: Elaboração própria

4.4.3 Exploração na rede de informação

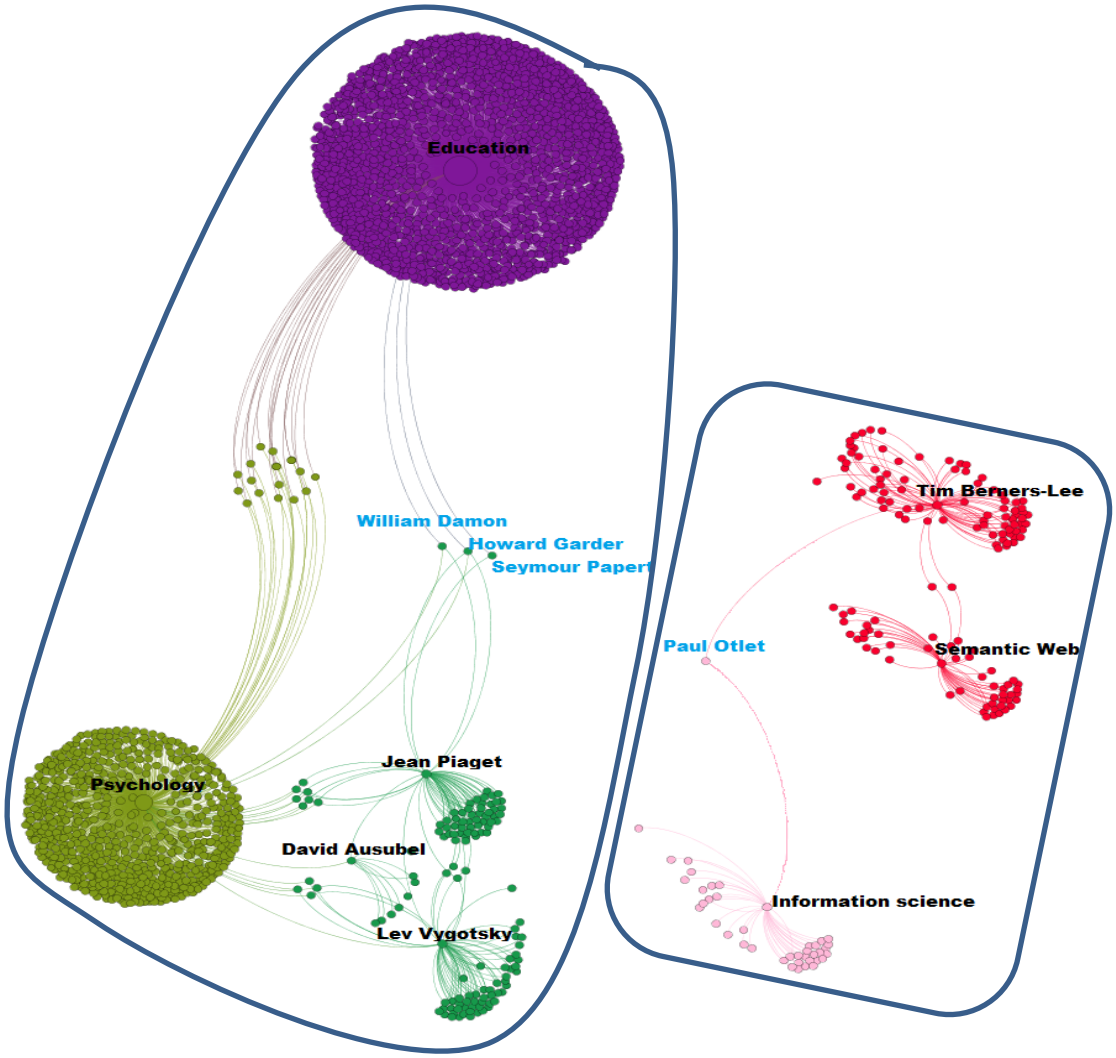
A exploração nas redes de informação intermediárias geradas ao longo do processo foi conduzida pelo método de inspeção visual e aplicação de operações de rede complexas com medidas, métricas e análise topológica.

A Figura 47 mostra a mesma rede da Figura 46, porém com formatações para auxiliar o processo de inspeção visual. Foi aplicado um *layout* pelo algoritmo Yifan Hu[76]. As cores diferentes dos nós indicam o resultado de um particionamento, realizado com o software Gephi, para auxiliar o estudo das relações entre os termos de consulta. Os cinco grupos coesos destacados pelo particionamento são: (1) ‘Education’, (2) ‘Psychology’, (3) ‘Jean Piaget’, ‘Lev Vygotsky’ e ‘David Ausubel’, (4) ‘Tim Berners-Lee’ e ‘Semantic web’, e (5) ‘Information Science’. Foram gerados dois componentes conectados, marcados de forma

[76] Algoritmo de distribuição de redes disponível no Gephi. Página do seu autor, Dr. Yifan Hu: <http://yifanhu.net/index.html>

manual na Figura 47. Cada um dos componentes possui nós que fazem intermediação entre grupos coesos, tais como os nomeados ‘William Damon’ no componente à esquerda, e ‘Paul Otlet’, no componente da direita. Foram destacados os nós com melhores valores de *betweenness* (discutido na subseção 2.5.4.6) com intuito de revelar aqueles que fazem intermediações entre as partições. Dentre esses nós, foram selecionados aqueles com maior proximidade dos outros nós, ou seja, maiores valores de *closeness* (discutido na subseção 2.5.4.7), e nomeados na cor azul.

Figura 47 – Rede de informação com formatações para inspeção visual

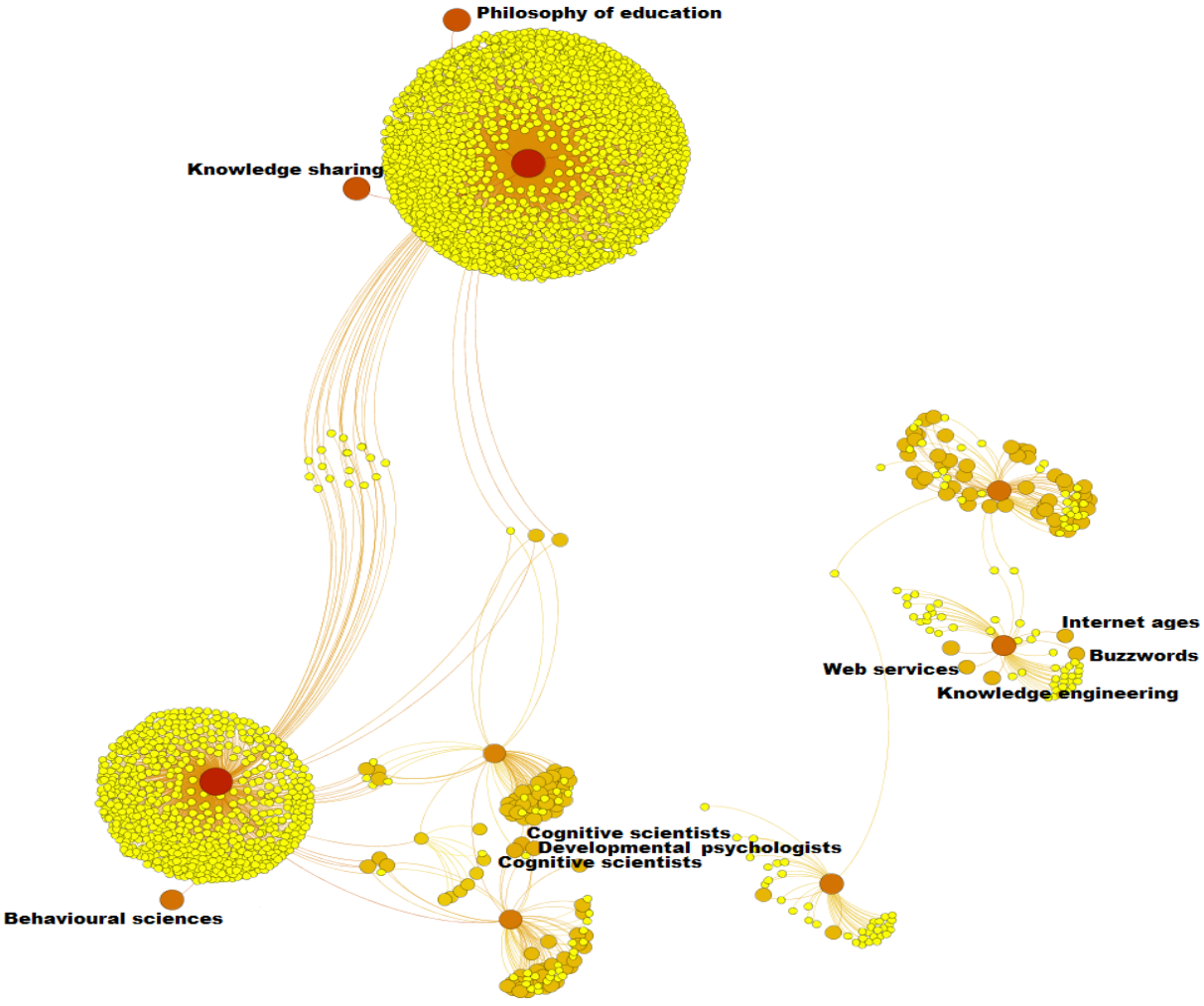


Fonte: Elaboração própria

Além dessa seleção, também foram identificados os nós que possuiam um bom grau de importância para os demais, por intermédio da métrica *eigenvector* (discutido na subseção 2.5.4.8). Seus nomes são mostrados na rede da Figura 48. Outras métricas que poderiam ser

usadas, mostraram-se inadequadas, por exemplo: (i) *in-dregree* (discutido na subseção 2.5.4.1): apesar de medir a quantidade de nós que fazem referência, não consegue capturar a importância deles; (ii) *PageRank*[77]: apesar de similar a *eigenvector*, há um fator de aleatoriedade em sua medida que poderia trazer alguma instabilidade na recuperação de informação; (iii) *out-degree* (discutido na subseção 2.5.4.1): não reflete integralmente a importância do nó pois revela apenas a sua dependência com aqueles que recebem a sua conexão.

Figura 48 – Rede de informação com destaque para maiores *eigenvector*



Fonte: Elaboração própria

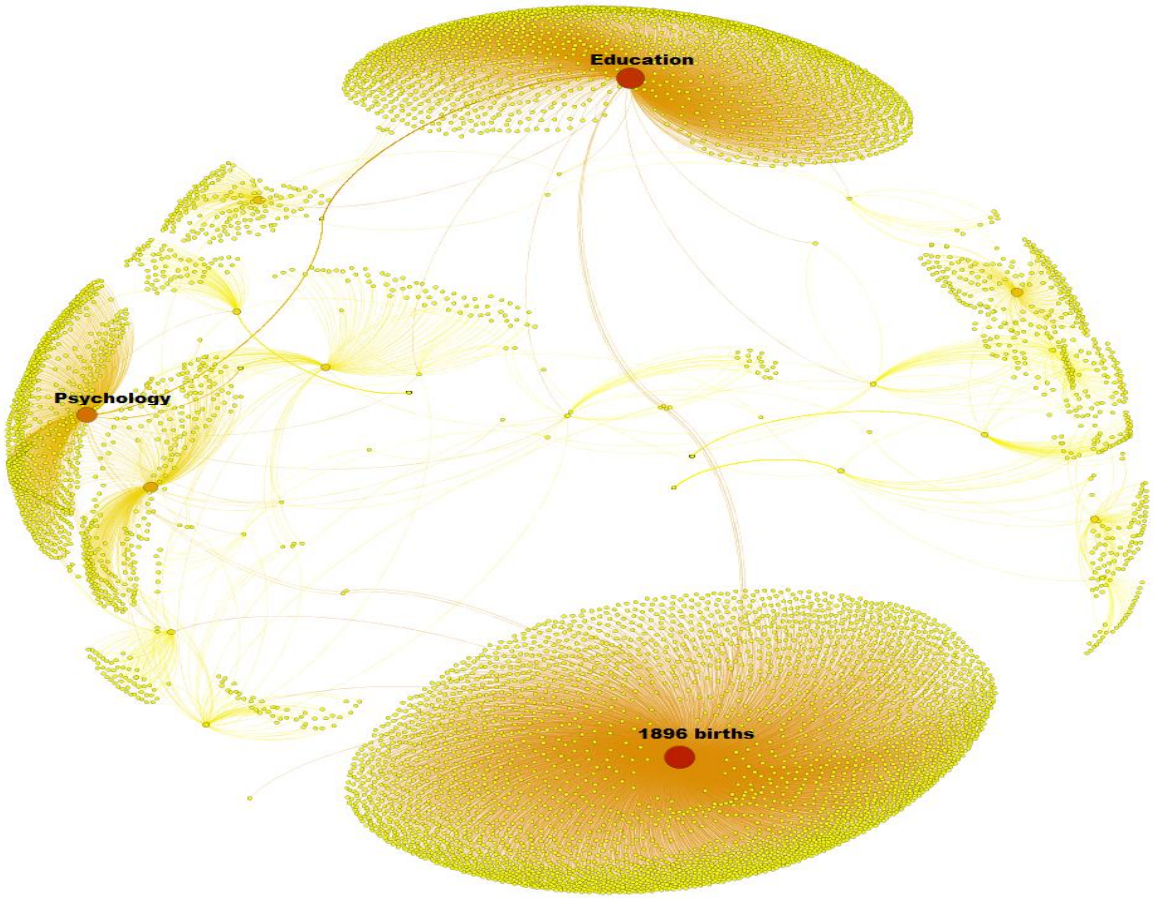
Foram então selecionados 14 termos, sendo 4 por *betweenness* e *closeness*, e 10 por eigenvector: ‘Howard Gardner’, ‘William Damon’, ‘Seymour Papert’, ‘Paul Otlet’,

[77] PageRank: conhecido algoritmo de busca usado pelo Google.

‘Behavioural sciences’, ‘Philosophy of education’, ‘Knowledge sharing’, ‘Cognitive scientists’, ‘1896 births’, ‘Developmental psychologists’, ‘Web services’, ‘Internet ages’, ‘Knowledge engineering’ e ‘Buzzwords’.

Como próxima fase do experimento, esses 14 nós selecionados serviram de base para nova consulta na base de conhecimento com o intuito de adicionar na rede novas triplas RDFs e, assim, tentar unificar os dois componentes conectados bem como achar outros nós importantes para o relacionamento dos termos de consulta do usuário. A Figura 49 mostra a rede com o acréscimo dos novos nós e conexões resultantes da consulta e mapeamento. Foram destacados os três nós com maior grau dentre o total de 7.210 nós e 7.383 conexões. O algoritmo de layout usado foi Yifan Hu e sem particionamentos.

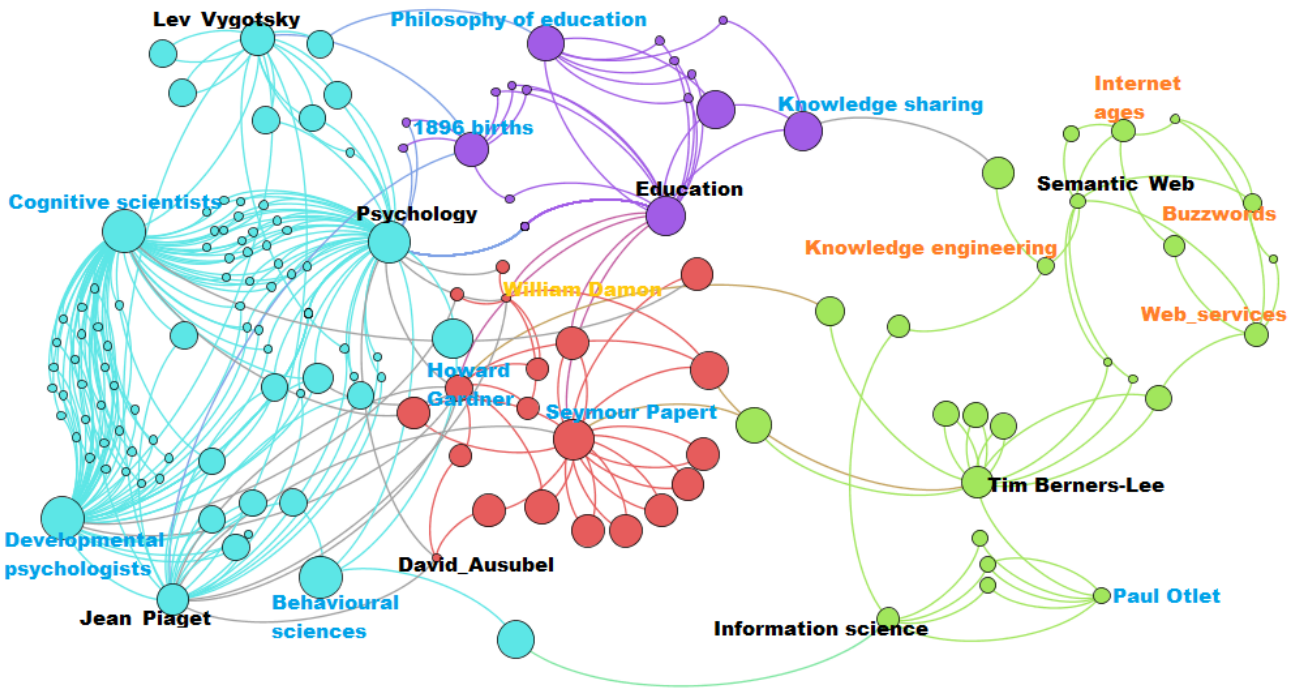
Figura 49 – Rede de informação formada pelas subredes dos 8 termos de consulta do usuário acrescidos dos 14 termos selecionados, destaque para os nós de maior grau



Fonte: Elaboração própria

Observou-se nessa rede que o nó denominado “1896 births” tinha valor de grau elevado, por fazer referência a um grande número personalidades que nasceram nesse ano, sendo que sua seleção anterior aconteceu pelo fato de dois nós estarem relacionados a ele (‘Jean Piaget’ e ‘Lev Vygotsky’ nasceram em 1896). Também foi observada que a grande maioria dos nós da rede era de grau 1. Isso poderia distorcer medidas que tentassem extrair o relacionamento entre os nós cuja importância fosse maior, como aqueles que representam os 8 termos de consulta do usuário e os 14 termos adicionados. Em função disso, optou-se por aplicar o algoritmo K-core (discutido na subseção 2.5.4.5) com índice 2. Dessa forma, várias medidas passaram a identificar com mais precisão os nós mais importantes entre aqueles já selecionados.

Figura 50 – Rede de informação 2-core, com destaque dos nós inseridos e excluídos



Fonte: Elaboração própria

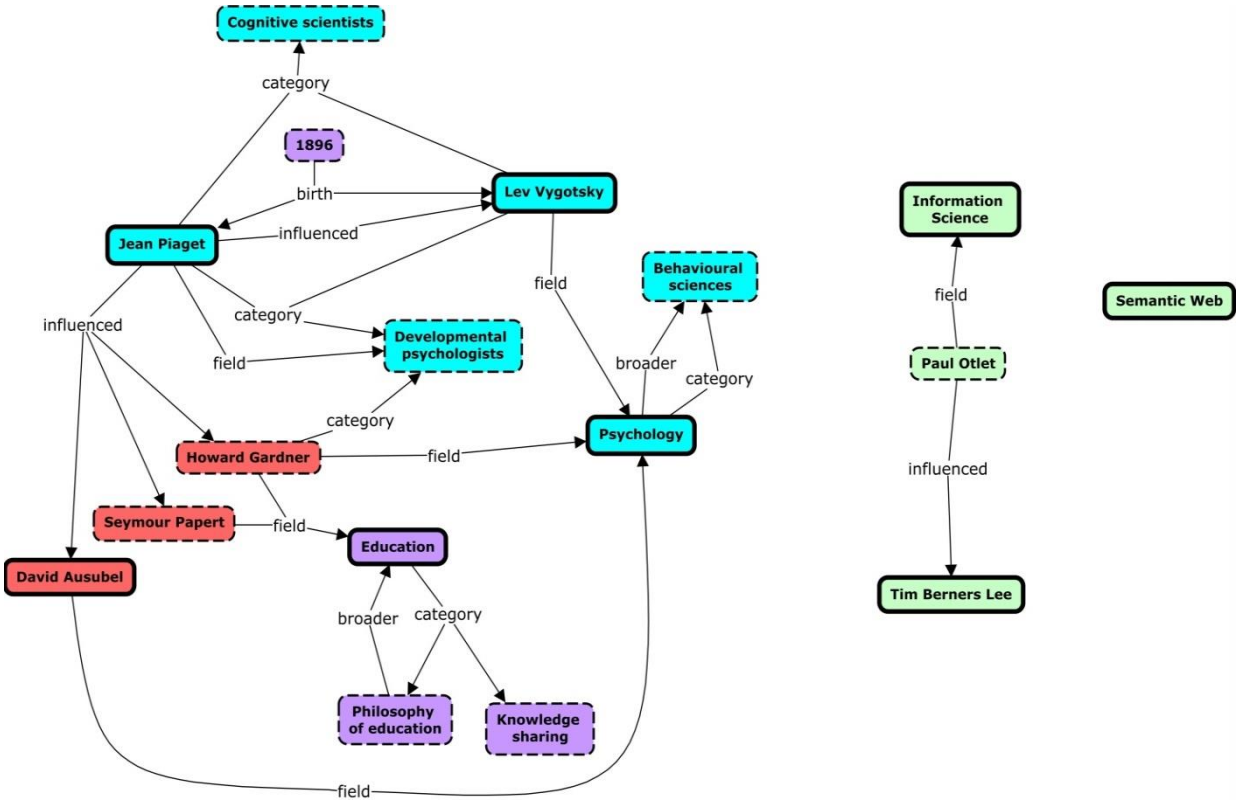
Após a aplicação de 2-core a rede passou a ter 169 nós e 342 conexões, mostrada na Figura 50. Com o intuito de diminuir a quantidade de nós para o último mapeamento e a criação do mapa conceitual resultante, foi feito um ranking sobre a métrica *eigenvector* naqueles 14 nós anteriormente selecionados. Permaneceram os 9 maiores, identificados com o nome em azul, e excluídos os 5 menores, identificados com o nome em laranja. Após a execução de particionamento para identificação de grupos coesos, foi possível inferir sobre a semântica dos grupos, por exemplo, o azul relacionado a temas da Psicologia e Ciências

Sociais, enquanto o vermelho relacionado a personalidades envolvidas com a educação, o roxo destaca elementos da Educação e, finalmente, o verde indica elementos próximos à Ciência da Informação.

4.4.4 Construção do mapa conceitual resultante

O software CmapTools auxiliou a construção do mapa conceitual, Figura 51, proveniente do ranqueamento realizado na ultima rede de informação, Figura 50. Foi usado um vocabulário controlado (uma versão simplificada do vocabulário mostrado no APÊNDICE F) para a transição da notação de predicados para a notação de frases de ligação. As cores do particionamento feito na rede de informação foram mantidas, os conceitos provenientes dos termos de consulta do usuário foram destacados com uma borda mais espessa e os conceitos adicionados com uma borda pontilhada.

Figura 51 - Mapa conceitual parcial, proveniente da rede de informação intermediária



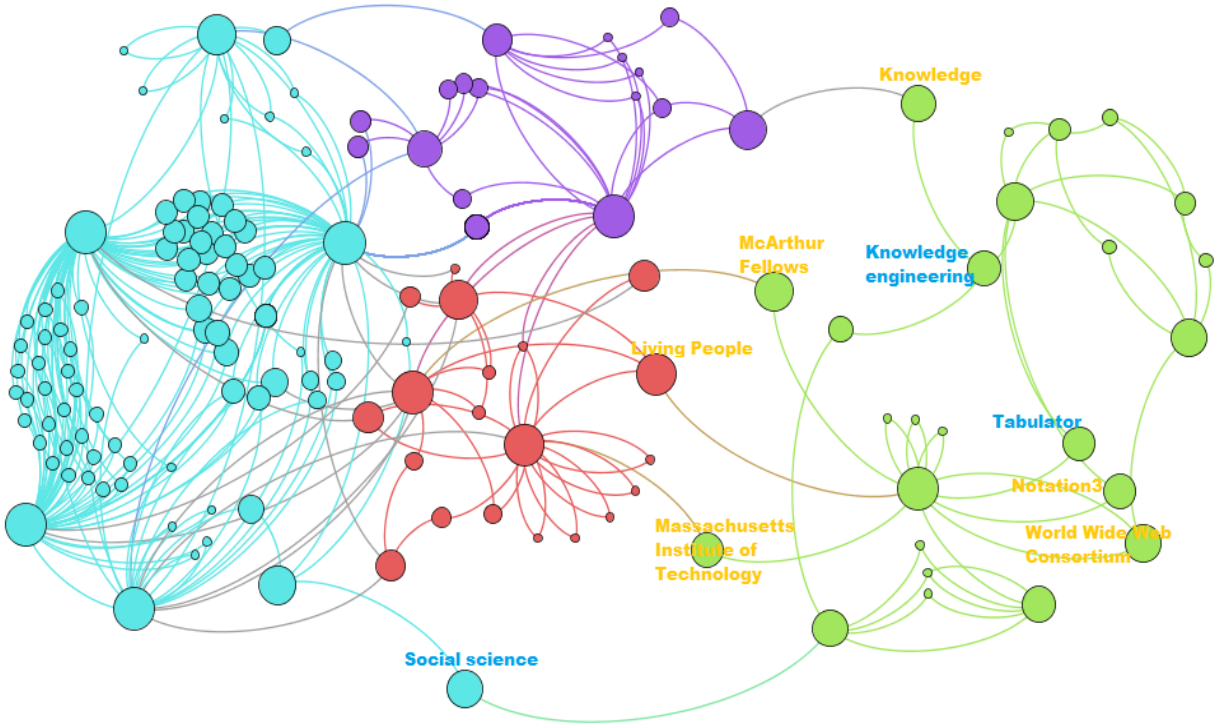
Fonte: Elaboração própria

Contudo, o mapa da Figura 51 possui conceitos sem ligação com os demais, sendo isso um problema para cumprimento do objetivo de estabelecer relações entre os termos de

consulta do usuário. Analisando o mapa como uma rede, observou-se a existência de 3 componentes conectados. Recorreu-se novamente à rede de informação anterior, Figura 50, para buscar elementos, pela métrica *betweenness*, que pudessem conectá-los, dando preferência aos nós que tinham conexão com os termos de consulta do usuário ou nós já selecionados. Dessa forma, a Figura 52 mostra a rede de informação com destaque em azul para os nós escolhidos ‘Knowledge engineering’ e ‘Tabulator’, e em laranja para os candidatos que foram descartados ‘Knowledge’, ‘Mc Fellows’, ‘Living People’ e ‘Massachusetts Institute of Technology’.

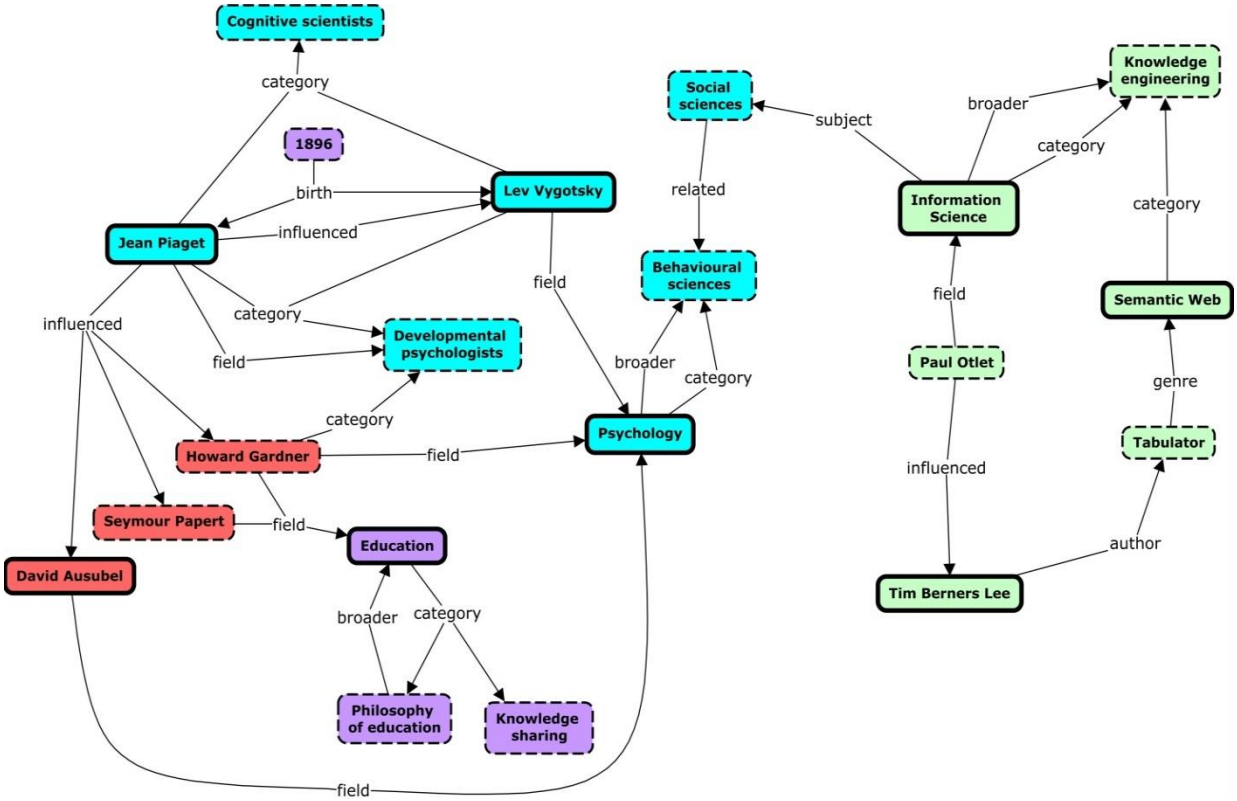
A Figura 53 apresenta o mapa conceitual final, proveniente da busca por ‘Education’, ‘Psychology’, ‘Jean Piaget’, ‘Lev Vygotsky’, ‘David Ausubel’, ‘Tim Berners-Lee’, ‘Semantic web’ e ‘Information Science’. Esse mapa contemla as novas proposições que uniram todos os conceitos num único componente conectado. Os três novos conceitos adicionados foram ‘Social sciences’, ‘Knowledge engineering’ e ‘Tabulator’.

Figura 52 – Rede de informação com destaque para os nós de intermediação, em azul os nós selecionados, e em laranja os nós descartados



Fonte: Elaboração própria

Figura 53 – Mapa conceitual resultante do experimento inicial



Fonte: Elaboração própria

4.4.5 Conclusões e problemas revelados no experimento inicial

A apresentação da informação recuperada em formato de mapa conceitual tem como um dos objetivos a revelação dos relacionamentos entre os termos inicialmente propostos, seja diretamente ou indiretamente. Isto é, não é interesse o descobrimento de atributos individuais de um ou outro termo, mas, apresentar conexões e novos termos que sejam importantes na intermediação dos termos originalmente escolhidos. A execução de um ciclo completo nesse experimento inicial em conjunto com uma análise exploratória da rede de informações permitiu verificar a factibilidade da continuidade da pesquisa.

Porém, alguns problemas foram detectados. Quanto ao conteúdo, observou-se que relacionamentos notoriamente conhecidos, tal como entre os conceitos ‘Tim Berners-Lee’ e ‘Semantic Web’ foi representado de forma pobre, nesse caso, com o conceito ‘Tabulator’, um navegador para web, ao invés de algo mais representativo como se conhece na literatura. Por conta disso, é possível que a repetição do processo de busca na base de conhecimento, com uma retroalimentação a partir de seleções de nós feita por intermédio de análise de redes, pudesse aumentar a rede intermediária e permitir a escolha de nós mais importantes para os

relacionamentos. Além disso, observou-se que apenas uma repetição poderia não ser suficiente para estabelecer relações entre todos os termos de consulta do usuário deixando vários componentes conectados no mapa resultante.

Outro problema detectado foi o processo muito lento para a realização de um ciclo completo, uma vez que quase todos os procedimentos foram manuais, apenas com auxílio de ferramentas isoladas de software. Em função dessa baixa velocidade, não houve possibilidade de validação com um grupo de usuários nesse estágio da pesquisa. Também em função da baixa velocidade, não foi possível experimentar outras métricas de rede ou a combinação delas.

4.5 Modelo aprimorado

Essa seção apresenta o modelo aprimorado em função dos problemas detectados no experimento e no modelo inicial da seção anterior (4.4), e da descoberta de novos elementos que serão apresentados nas próximas subseções. É também apresentado o algoritmo usado no protótipo implementado, a sua execução e alguns testes piloto realizados numa base de dados ligados privada e na DBpedia.

4.5.1 Aprimoramentos no modelo

A construção de um protótipo para automatizar quase por completo o processo, desde a montagem das consultas a partir dos termos do usuário até a exibição final do mapa conceitual resultante, foi determinante para tornar possíveis várias ações fundamentais. Principalmente pela velocidade maior, uma vez que a execução do ciclo completo que, na primeira versão, poderia demorar alguns dias, no modelo aprimorado passou a consumir desde o tempo de 1 minuto até cerca de 40 minutos, dependendo do tamanho da rede de informação intermediária. Algumas ações que foram possíveis com o protótipo e estão descritas a seguir.

- A descoberta de outros elementos de análise de redes para compor o algoritmo e melhorar a montagem da rede informacional intermediária, os ranqueamentos realizados e a seleção de elementos relevantes para compor o mapa conceitual resultante.
- A possibilidade de incluir iterações no algoritmo, permitindo uma retroalimentação sucessiva em várias passagens, ou seja, permitir que os elementos

melhores ranqueados na rede de informação fossem os termos para a próxima busca na base de conhecimento e assim sucessivamente.

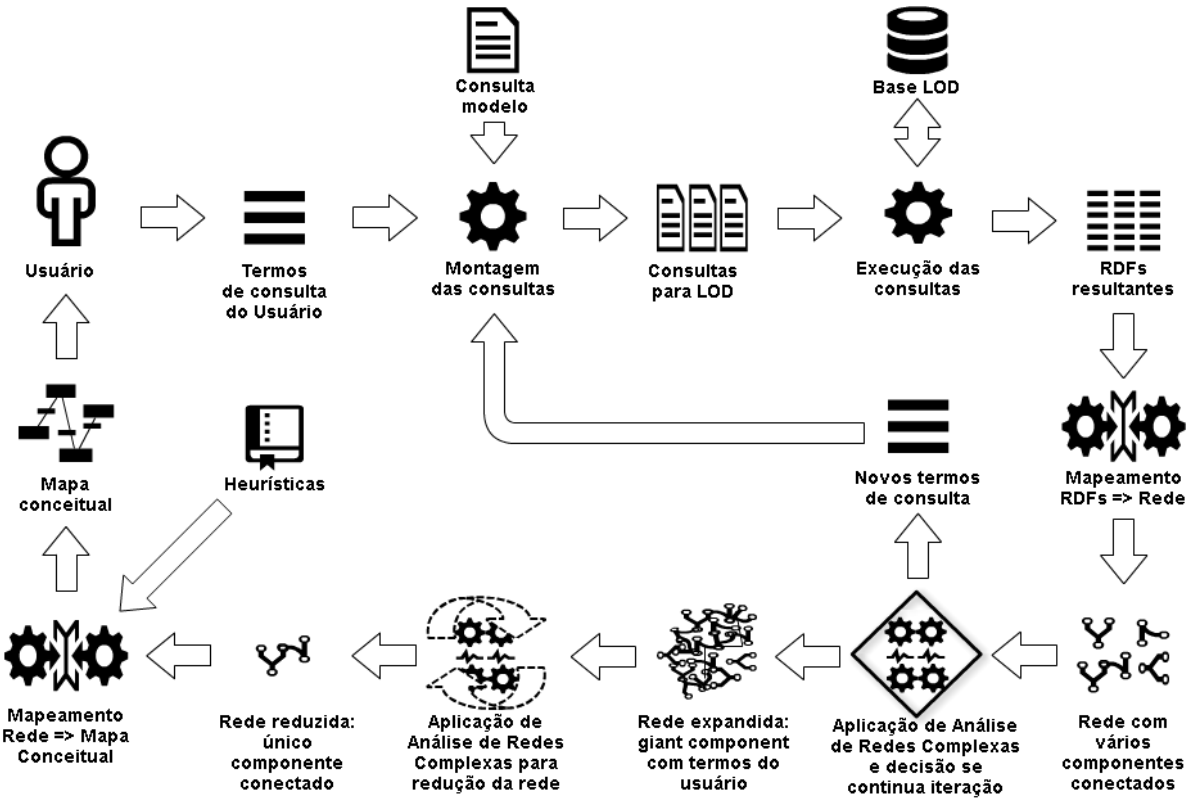
- A inclusão no algoritmo de um processo mais elaborado de redução da rede de informação, onde a rede é reconstruída a partir dos nós já secionados mantendo-se o caminho mínimo entre deles.
- O tratamento independente dos vários componentes conectados para o cálculo das métricas de rede.
- A flexibilização da quantidade de termos fornecidos pelo usuário, passando a aceitar qualquer quantidade maior ou igual a dois.
- O aumento da quantidade de testes para viabilizar comparações e análises comparativas entre várias situações e contextos informacionais de informação recuperada.
- O aumento da alterabilidade para permitir a configuração rápida do algoritmo, possibilitando experimentar situações diversas, tais como aumentar o peso da métrica *betweenness* em detrimento de *eigenvector* nos ranqueamentos, estipular a quantidade de conceitos no mapa resultante, determinar com facilidade a base de conhecimento a ser usada etc. O arquivo completo de configuração encontra-se no APÊNDICE E.
- A inclusão de heurísticas para melhoria da legibilidade do mapa conceitual resultante.
- A validação do modelo proposto com um grupo de usuários.
- A utilização de acesso direto a base de conhecimento DBpedia por meio do protótipo ao invés do uso indireto e manual por um terminal, como o SNORQL.

4.5.2 Diagrama geral

A Figura 54 apresenta o diagrama geral do modelo aprimorado. O usuário fornece um conjunto de termos para a busca que, baseado numa consulta modelo, são reescritos como consultas para *linked open data* (LOD). A execução dessas consultas sobre a base LOD recupera um conjunto de triplas RDFs resultantes, que passam por um mapeamento transformando-se numa rede de informação, normalmente com vários componentes conectados devido à distância semântica dos termos. É aplicada uma análise de redes complexas sobre a rede de informação para ranquear e selecionar nós em potencial, ou novos termos, para permitir a unificação dos vários componentes conectados. Se a rede ainda não possui um componente gigante que integre todos os termos do usuário, o fluxo do modelo

retorna retroalimentando uma nova busca com os nós selecionados, tendo os novos RDFs recuperados e mesclados na rede existente. Esse processo se repete enquanto o critério de unificação dos termos de consulta do usuário não for atendido. Por outro lado, continuando o fluxo, inicia-se o processo de redução da rede expandida que é feito por uma nova análise de redes com a construção de nova rede contendo os nós selecionados nas iterações anteriores, mantendo os caminhos mínimos entre os termos de consulta do usuário em um componente gigante. Finalmente, o mapeamento da rede de informação final é feito para o mapa conceitual resultante, tendo o auxílio de um vocabulário controlado e aplicação de algumas heurísticas para aumento da legibilidade do mapa. O detalhamento completo de todo esse processo é apresentado nas duas próximas subseções, que apresentam o diagrama de processos e o algoritmo.

Figura 54 – Diagrama geral do modelo aprimorado

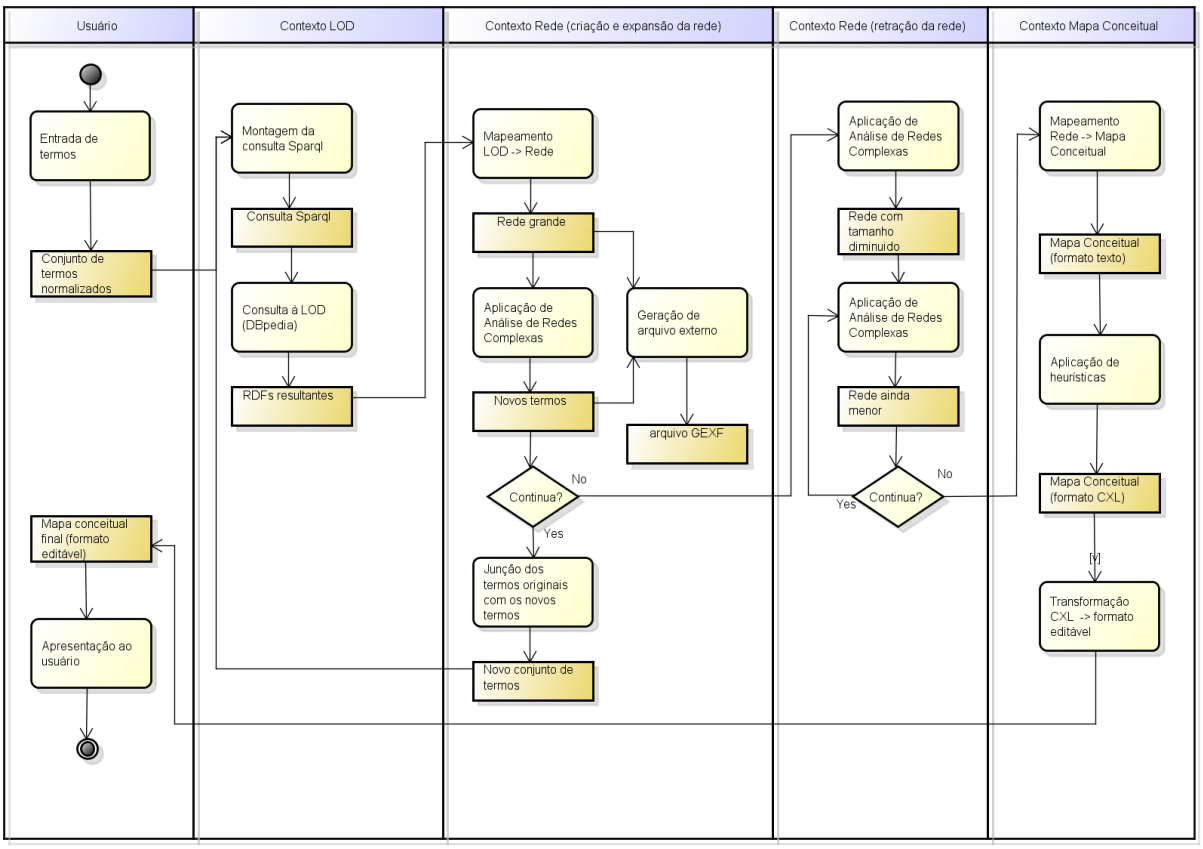


Fonte: Elaboração própria

4.5.3 Diagrama de processos

A Figura 55 contém o diagrama de processos do modelo aprimorado. O fluxo processual passa por cinco contextos, representados respectivamente pelas cinco colunas do diagrama, a saber: (i) **usuário**, desde o fornecimento dos termos até a sua normalização; (ii) **contexto LOD**, desde a montagem da consulta SPARQL, consulta na base de conhecimento formada por dados abertos ligados até a obtenção das triplas RDFs; (iii) **contexto rede (criação e expansão da rede)**, começando pela montagem de rede de informação, aplicação de análise de rede, geração do arquivo externo GEXF[78] para facilitar a inspeção visual com o software Gephi, e a verificação de continuidade da iteração para expansão da rede; (iv) **contexto rede (retração da rede)**, com a aplicação de análise de redes para exclusão de nós e conexões até atingir o tamanho previsto; (v) **contexto mapa conceitual**, com o mapeamento para montagem do mapa resultante, aplicação de heurísticas, até a transformação do mapa num formato editável, pelo software CmapTools, e entregue ao usuário.

Figura 55 – Diagrama de processos do modelo aprimorado



Fonte: Elaboração própria

[78] GEXF: *graph Exchange XML Format*, é uma linguagem de descrição de estruturas de redes complexas.

O diagrama de processos anotado, com informações extras a respeito de alguns procedimentos, encontra-se no APÊNDICE B.

4.5.4 Algoritmo

O detalhamento do modelo aprimorado (Figura 54) e do diagrama de processos (Figura 55) pode ser verificado no algoritmo do Quadro 7. Alguns passos desse algoritmo são explicados ou justificados nos próximos tópicos:

- Passo (1): os termos do usuário são normalizados, isto é, passam por adequação quanto à abreviação, uso de letras maiúsculas e minúsculas, uso de caracteres especiais, tais como acentuação, hífen etc.
- Passo (2): a busca de RDFs na base de dados ligados acontece por intermédio de uma consulta modelo (ver APÊNDICE D), onde são configuradas características desejadas para o resultado, tais como filtros para seleção de partes específicas da base, opção por língua entre outros.
- Passos (3) e (4.i): o mapeamento de RDFs para a rede acontece através de 11 casos distintos, considerando repetições entre *subject*, *predicate* e *object* (ver subseção 4.3.1).
- Passo (4.b): o corte de nós com grau 1 é devido à necessidade de diminuição do tempo de cálculo das métricas que aumentam exponencialmente em função da quantidade de nós na rede. Esse filtro só é aplicado se houver apenas um componente conectado, pois caso contrário poder-se-ia prejudicar possíveis conexões entre componentes ainda não conectados, já que as ‘pontas’ da rede (nós com grau um) desses componentes seriam cortadas.
- Passo (4.c): uma das principais metas do algoritmo é achar relacionamentos entre os termos do usuário, seja por meio de ligações diretas ou indiretas passando por outros conceitos. Sendo assim, o laço (4) das iterações que retroalimentam a rede com novos RDFs deve continuar iteragindo enquanto a rede não estiver totalmente conectada (1 componente conectado) ou, pelo menos, existam caminhos possíveis entre os termos do usuário.
- Passo (4.d): a justificativa para o uso das métricas *betweenness*, *closeness* e *eigenvector* é a mesma já discutida na subseção 4.4.3 sobre o experimento inicial.

Quadro 7 – Algoritmo do modelo aprimorado

1. Entrada dos termos do usuário e normalização para aceitação na base de conhecimento.
2. Busca, na base LOD, dos RDFs referentes a cada termo do usuário.
3. Mapeamento dos RDFs coletados no passo (2) para uma rede informacional.
4. LAÇO das iterações que retroalimentam a rede com novos RDFs.
   a) Cálculo quantidade de componentes conectados na rede e quantidade de caminhos entre os termos do usuário.
   b) Se existir apenas um componente conectado na rede, a partir de uma determinada iteração no laço ENTÃO remove todos os nós da rede com grau igual a 1.
   c) Se atingir um determinado limite mínimo de iterações e a quantidade de componentes conectados é igual a 1 OU
   se atingir um determinado limite mínimo de iterações e existem conexões entre todos os termos do usuário
   ENTÃO sai do laço (4) e vai para o passo (5).
   d) Calculo das métricas betweenness, closeness e eigenvector para cada nó da rede.
   e) Ordenação decrescente por betweenness para todos os nós da rede e, entre os primeiros, ordenação decrescente por closeness.
   f) Ordenação decrescente por eigenvector para todos os nós da rede.
   g) Seleção dos primeiros nós de cada ordenação em (4.e) e (4.f), separadamente para cada componente conectado, excluindo-se aqueles selecionados em iterações anteriores.
   h) Busca na base de dados ligados os RDFs referentes aos nós da seleção em (4.g).
   i) Mapeamento e união dos novos RDFs coletados em (4.h) na rede.
   j) Retorno ao passo (4.a).
5. Se quantidade de componentes conectados é igual a 1
ENTÃO aplique K-core na rede resultante (por default K=2).
6. Reunião de todos os nós que foram selecionados no passo (4.d) pelo critério de maior betweenness, porém com a exclusão daqueles cujo betweenness seja igual a zero na última iteração. Acréscimo dos nós referentes aos termos do usuário.
7. Determinação de todos os caminhos mínimos entre todos os nós selecionados no passo anterior (6) e a seleção dos nós pertencentes a esses caminhos.
8. Criação de uma rede contendo apenas os nós, com suas respectivas arestas, selecionados no passo anterior (7).
9. Cálculo da quantidade ideal de conceitos resultantes no mapa conceitual.
10. LAÇO das iterações que diminuem nó a nó o tamanho da rede.
   a) Cálculo da métrica eccentricity pra todos os nós e ordenação crescente em função desse valor.
   b) Seleção do nó com menor eccentricity no conjunto ordenado.
   c) Exclusão do nó selecionado em (10.b).
   d Cálculo da quantidade de componentes conectados.
   e) Se a quantidade de componentes conectados aumentar.
   ENTAO recupera o nó recém excluído, avança uma posição no conjunto ordenado por eccentricity e volta ao passo (10.b).
   f) Se a quantidade de nós da rede for igual a quantidade ideal para o mapa conceitual resultante OU se não existem mais nós para tentar excluir
   ENTÃO sai do laço (12) e vai para o passo (11).
11. Mapeamento da rede resultante num mapa conceitual.
12. Aplicação de heurísticas para melhoria da legibilidade do mapa conceitual gerado em (11).
13. Aplicação de heurística, específica para o contexto da base de conhecimento usada, com o objetivo de obter melhor distribuição entre tipos de conceitos resultantes.

Fonte: Elaboração própria

- Passos (4.e), (4.f) e (4.g): a seleção dos nós com melhores *betweenness+closeness* e *eigenvector* foi feita em cada componente conectado na mesma quantidade, e de forma proporcional a quantidade de nós. Isto para permitir o crescimento da rede em cada componente de forma independente e assim aumentar a possibilidade de junção entre os nós de diferentes componentes, diminuindo a quantidade de componentes conectados. A seleção dos melhores *betweenness+closeness* ocorre primeiro pela seleção dos melhores valores de *betweenness* e depois, dentro dessa seleção, escolhem-se os de maior valor considerando-se a ordenação por *closeness*. Mesmo os nós com *betweenness* zerado podem ser escolhidos em função da possibilidade de existência de vários componentes conectados e, nesse caso, ocorre a inexistência de nós com *betweenness* maiores que zero, exceto os que representam os termos do usuário, que não podem ser escolhidos.
- Passo (5): se o K-core (k=2) fosse aplicado prematuramente, no interior do laço (4), muitos nós com boa possibilidade de promover novas conexões seriam removidos. Portanto ele foi deixado para um ponto posterior ao laço (4) a fim de cumprir o objetivo de diminuir o tamanho da rede que foi bastante aumentado nas iterações desse laço. Devido ao aumento do custo computacional, foi descartada uma possível modificação no algoritmo K-core para que ele preservasse as conexões entre os termos do usuário. Porém, constatou-se, empiricamente, que a execução do K-core não interfere significativamente no *giant componente* formado até o momento.
- Passo (6): a decisão de resgatar todos os nós selecionados por meio de *betweenness+closeness*, em todas as iterações, foi em função da necessidade de se obter nós mais representativos no quesito importância de ligação entre as partes da rede. Se a medida *eigenvector* fosse usada nessa fase o resultado poderia tender para nós com mais importância por si só, e não em nós cuja importância fosse baseada no relacionamento entre as várias partes da rede. O acréscimo da exigência de *betweenness* maior que zero é devido ao fato de que elementos em componentes conectados muito pequenos tem esse valor zerado.
- Passos (7) e (8): com o objetivo de diminuir mais rapidamente o tamanho da rede, essa seleção é realizada. Ela garante que os nós que representam os termos do usuário bem como todos aqueles que servem de ligação intermediária entre eles estejam presentes na nova rede.

- Passo (9): por intermédio de uma função logarítmica que tem como entrada a quantidade de termos do usuário, o objetivo desse cálculo é que a quantidade final de conceitos no mapa não sofra grande variação caso a quantidade de termos do usuário cresça ou diminua muito. Dessa forma, qualquer mapa conceitual resultante fica sempre limitado a uma faixa restrita de número de conceitos, preservando a sua legibilidade.
- Passo (10.a) e (10.b): *eccentricity* é usada para selecionar o nó mais afastado do centro da rede como candidato para exclusão e, assim, tentar com uma boa margem de acerto, a preservação da quantidade de componentes conectados.
- Passo (10.e): se a exclusão do nó selecionado em (10.b) aumentar a quantidade de componentes conectados então aumenta-se muito a possibilidade do mapa conceitual não possuir seus conceitos totalmente conectados. Nesse caso recupera-se o nó para a rede e pega-se o próximo nó da lista ordenada por *eccentricity* para, em seguida, tentar uma nova exclusão e assim sucessivamente.
- Passo (10.f): o laço termina se a quantidade ideal de nós for atingida ou se não existirem mais nós para tentativa de exclusão. Nesse último caso o mapa conceitual ficará maior do que a meta, porém, em compensação, conserva-se a ligação dos conceitos.
- Passo (11): o mapeamento da rede num mapa conceitual é realizado por intermédio de 11 casos distintos de mapeamento (ver subseção 4.3.1). A questão focal do mapa é sempre definida como “Qual é o relacionamento entre (os termos do usuário)?”.
- Passo (12): as heurísticas aplicadas no mapa conceitual resultante são baseadas em conhecimento de especialistas na construção de mapas conceituais para obtenção de melhoria na legibilidade: (i) aplicação de vocabulário para transformar descrições de predicados de RDFs em frases de ligação nas proposições do mapa conceitual, (ii) limpeza de anotações realizadas nos links na fase do mapeamento , (iii) união de conceitos com label parecido, com o termo ‘categoria’ e sem esse termo, (iv) eliminação de ligações para o mesmo conceito, (v) transformação do mapa conceitual para o formato do CmapTools para a formatação de cores nos conceitos a fim de distinguir os conceitos fornecidos pelo usuário dos conceitos novos sugeridos pelo algoritmo, (vi) inserção de resumo ou comentários descritivos sobre os conceitos, coletados na fase de leitura dos RDFs.

- Passo (13): para evitar a predominância demasiada de um tipo de conceito no mapa conceitual, tal como pessoas, empresas, instituições e categorias de classificação, aplica-se essa heurística com o intuito de se atingir uma distribuição mais homogênea entre os tipos. Isso é conseguido com a alteração das proporções de *betweenness* e *eigenvector* sobre a quantidade de conceitos a serem buscados na base de conhecimento a cada iteração (laço 4) e também sobre os conceitos selecionados para composição dos caminhos mínimos na parte da retração da rede (laço 10).

4.5.5 Execução do protótipo

Os métodos empregados na construção do protótipo estão descritos no capítulo da metodologia, subseção 3.4, inclusive a indicação do repositório público onde está armazenado o código fonte. O protótipo implementado não representa a totalidade do modelo aprimorado, pois não foram implementadas as três seguintes partes: (i) a interface com o usuário que cuida da entrada dos termos de consulta, (ii) a heurística de configuração do peso das métricas (discutido na subseção 5.5.2) para distribuir uniformemente a quantidade de tipos de conceitos, tal como pessoas, empresas, instituições e categorias de classificação, e (iii) o layout automático final do mapa conceitual resultante, feito pelo software CmapTools, mas, que ainda necessita de alguns ajustes manuais para melhorar a legibilidade.

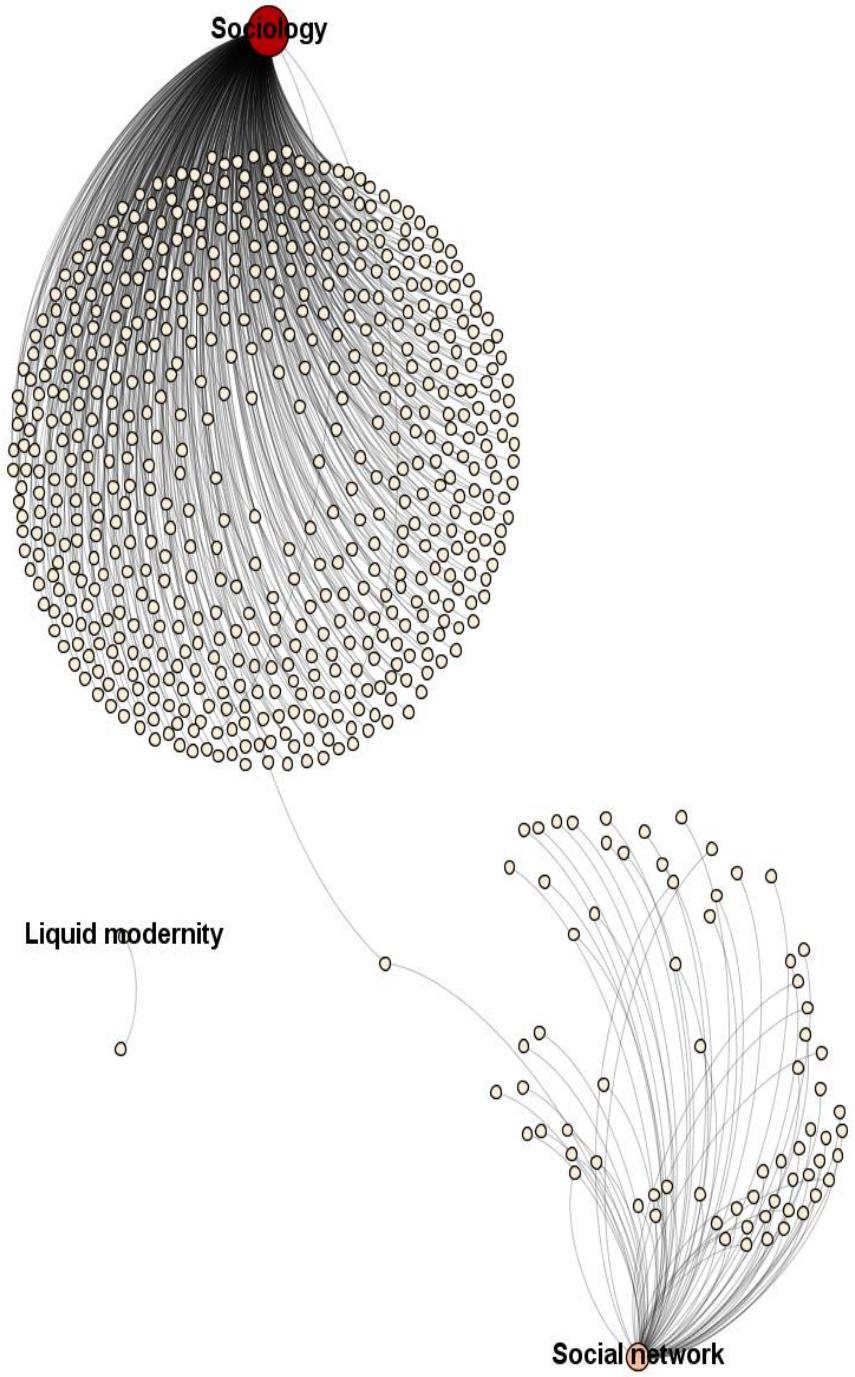
A configuração para a execução é feita em um arquivo específico, disponível no APÊNDICE E, onde são determinadas informações sobre o local da base de conhecimento, arquivos gerados, ajustes no algoritmo entre vários outros parâmetros. Durante a execução do protótipo são gerados arquivos de *log* que mostram o passo-a-passo detalhado de todas as operações realizadas bem como as redes de informação intermediárias, para cada iteração, geradas em formato legível pelo software Gephi.

A Figura 56 até a Figura 65 mostram estágios intermediários do processamento de uma solicitação de usuário que participou da validação do modelo e que representa uma necessidade informacional sobre a busca de relacionamentos entre os termos ‘Sociology’, ‘Liquid modernity’ e ‘Social network’. As redes de informação estão com formatação para destacar os nós com maior betweenness. O Quadro 8 contém o log resumido desse processamento.

Com apoio da visualização das redes de informação é possível observar mais facilmente estados intermediários e passagens do algoritmo. As seis primeiras figuras, Figura 56 a Figura 61, mostram o crescimento da rede de informação atingindo o seu ápice de 6.023

nós (log do Quadro 8 ‘Iteration 5’) em seguida, as três seguintes figuras, Figura 61 a Figura 64, mostram o processo de redução da rede até atingir a quantidade de 9 nós e, finalmente, a Figura 64 apresenta o mapa conceitual resultante com esses 9 nós mapeados em 9 conceitos.

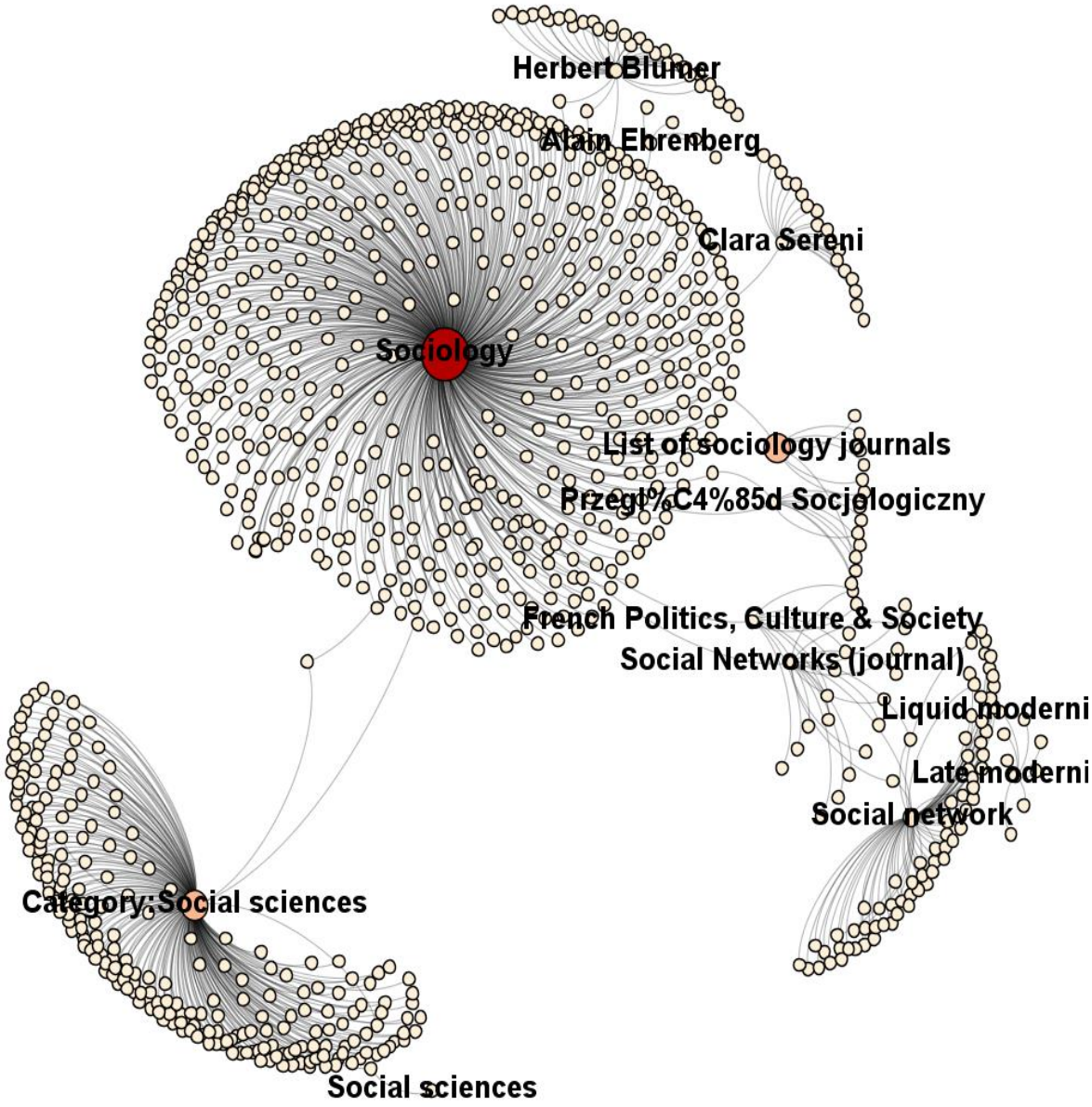
Figura 56 – Rede de informação referente à primeira iteração do algoritmo



Fonte: Elaboração própria, por intermédio do protótipo e apoio do software Gephi

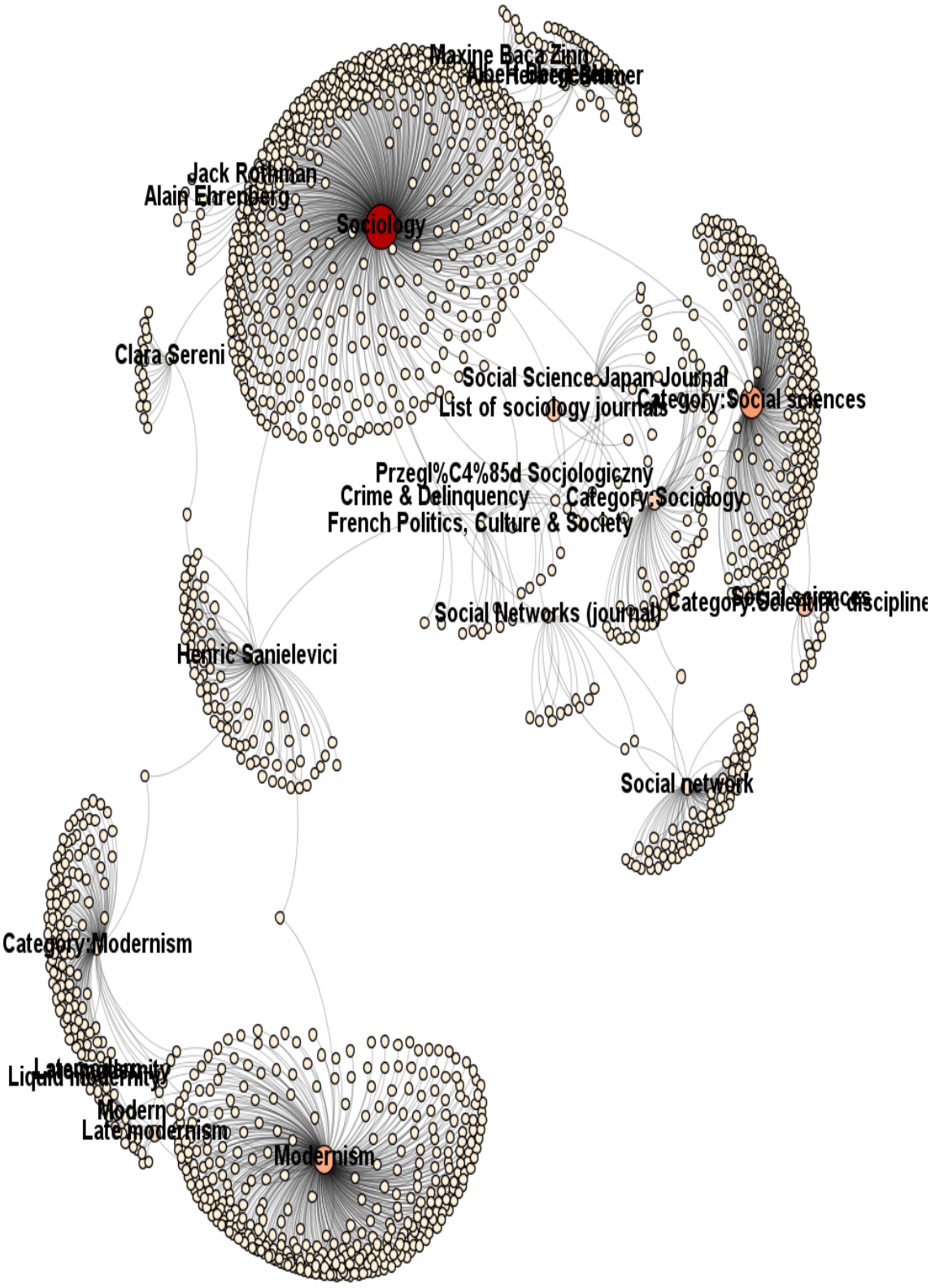
Na primeira rede de informação, Figura 56, dois termos do usuário, ‘Sociology’ e ‘Social network’, conseguiram conexão entre si, e o nó proveniente do termo ‘Liquid modernity’ encontra-se em um componente conectado isolado dos demais. Somente na terceira iteração, Figura 58, é que ele consegue conexão com os demais nós formando único componente conectado, sendo mostrado no log do Quadro 8 como ‘*** Iteration 2 ... Connected components: 1’, e tendo a rede crescido de 657 nós para 1.692 nós.

Figura 57 – Rede de informação referente à segunda iteração do algoritmo



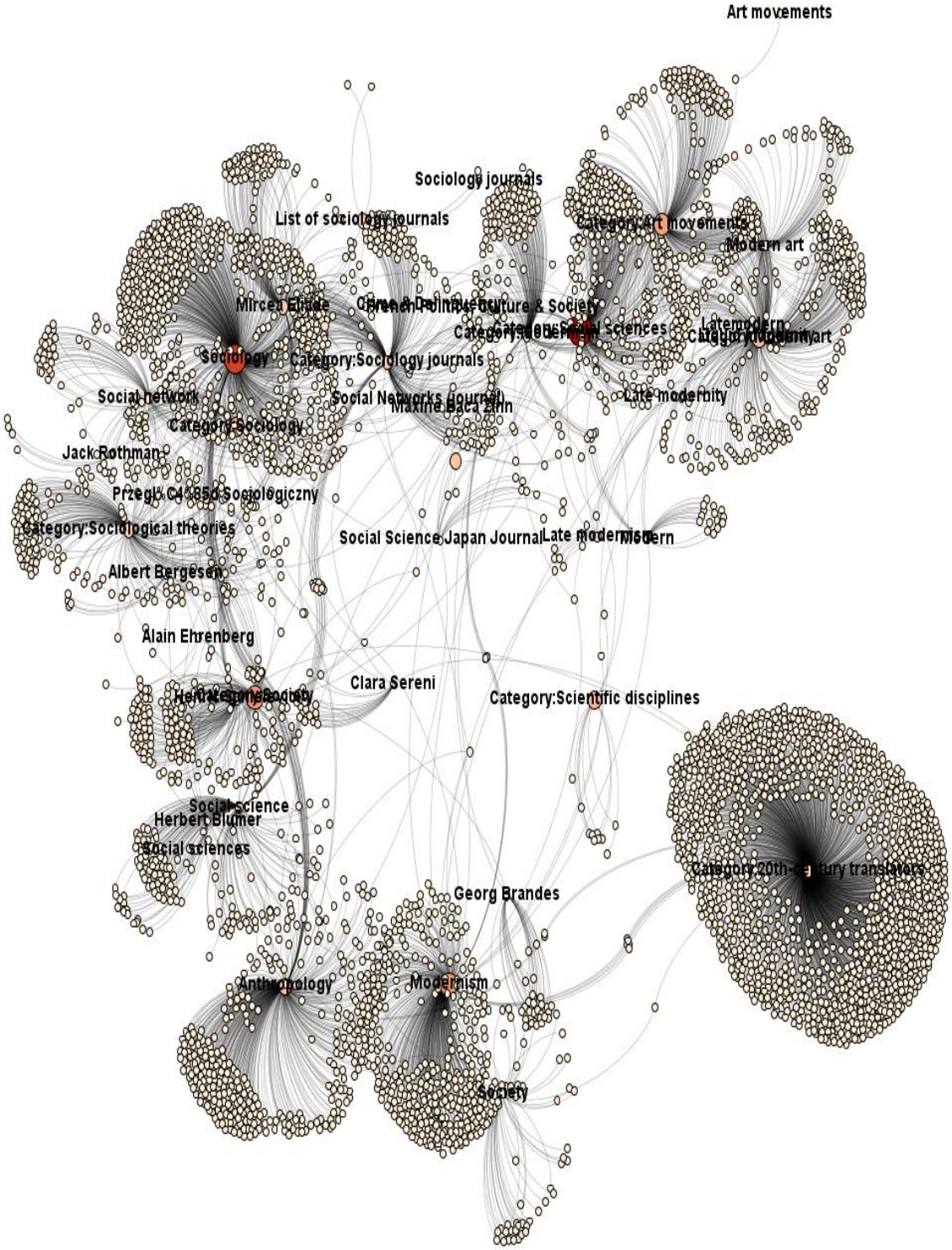
Fonte: Elaboração própria, por intermédio do protótipo e apoio do software Gephi

Figura 58 – Rede de informação referente à terceira iteração do algoritmo



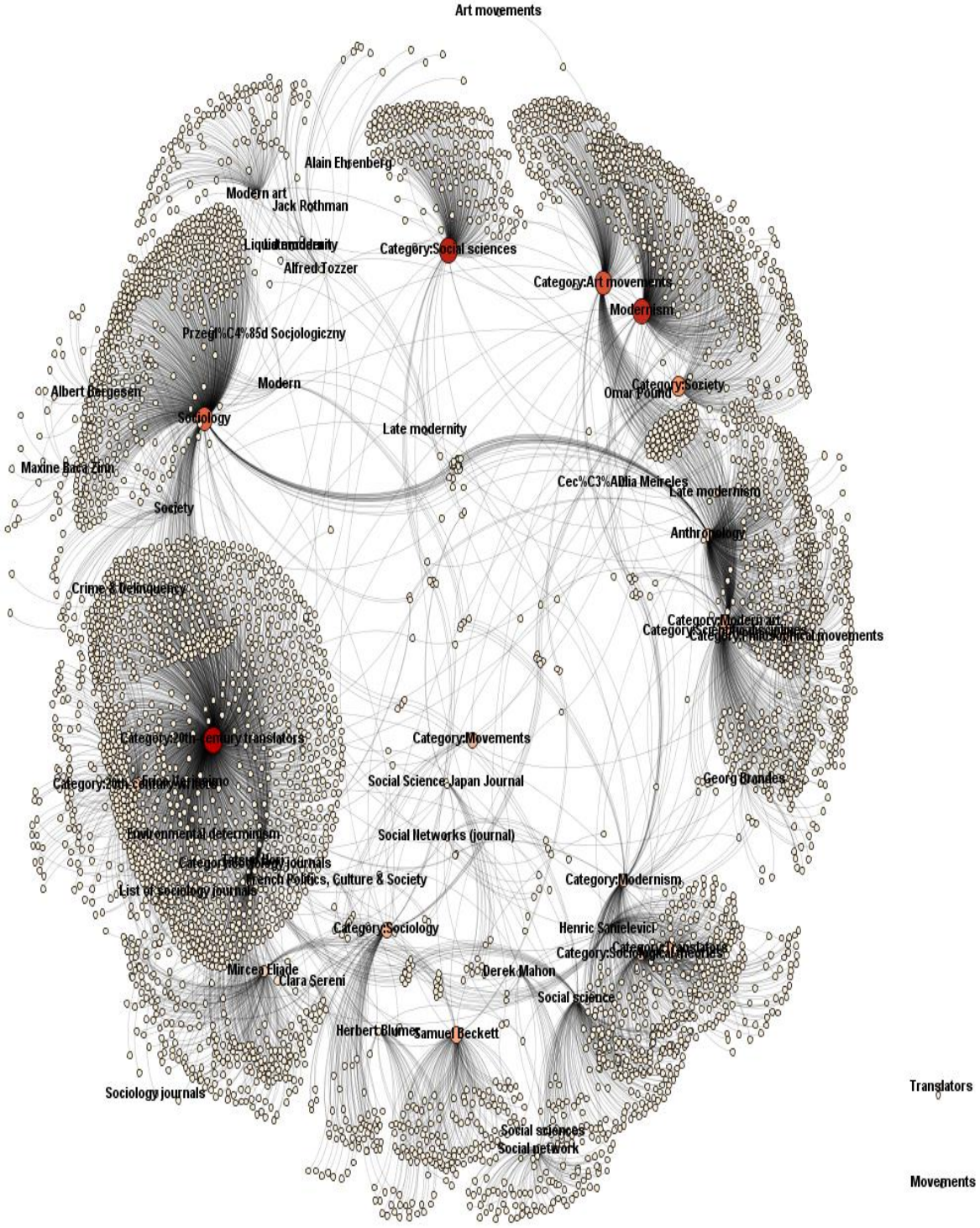
Fonte: Elaboração própria, por intermédio do protótipo e apoio do software Gephi

Figura 59 – Rede de informação referente à quarta iteração do algoritmo



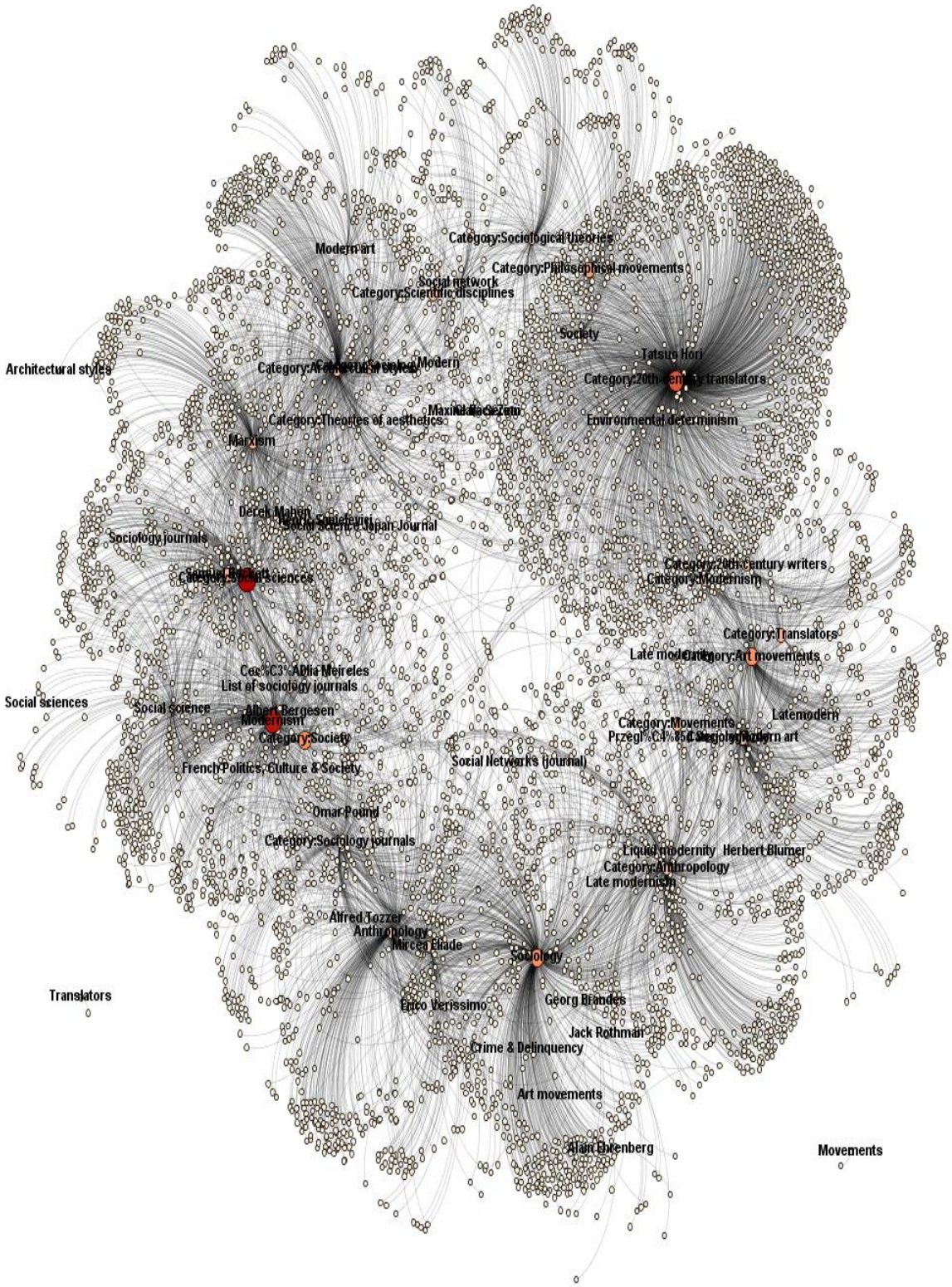
Fonte: Elaboração própria, por intermédio do protótipo e apoio do software Gephi

Figura 60 – Rede de informação referente à quinta iteração do algoritmo



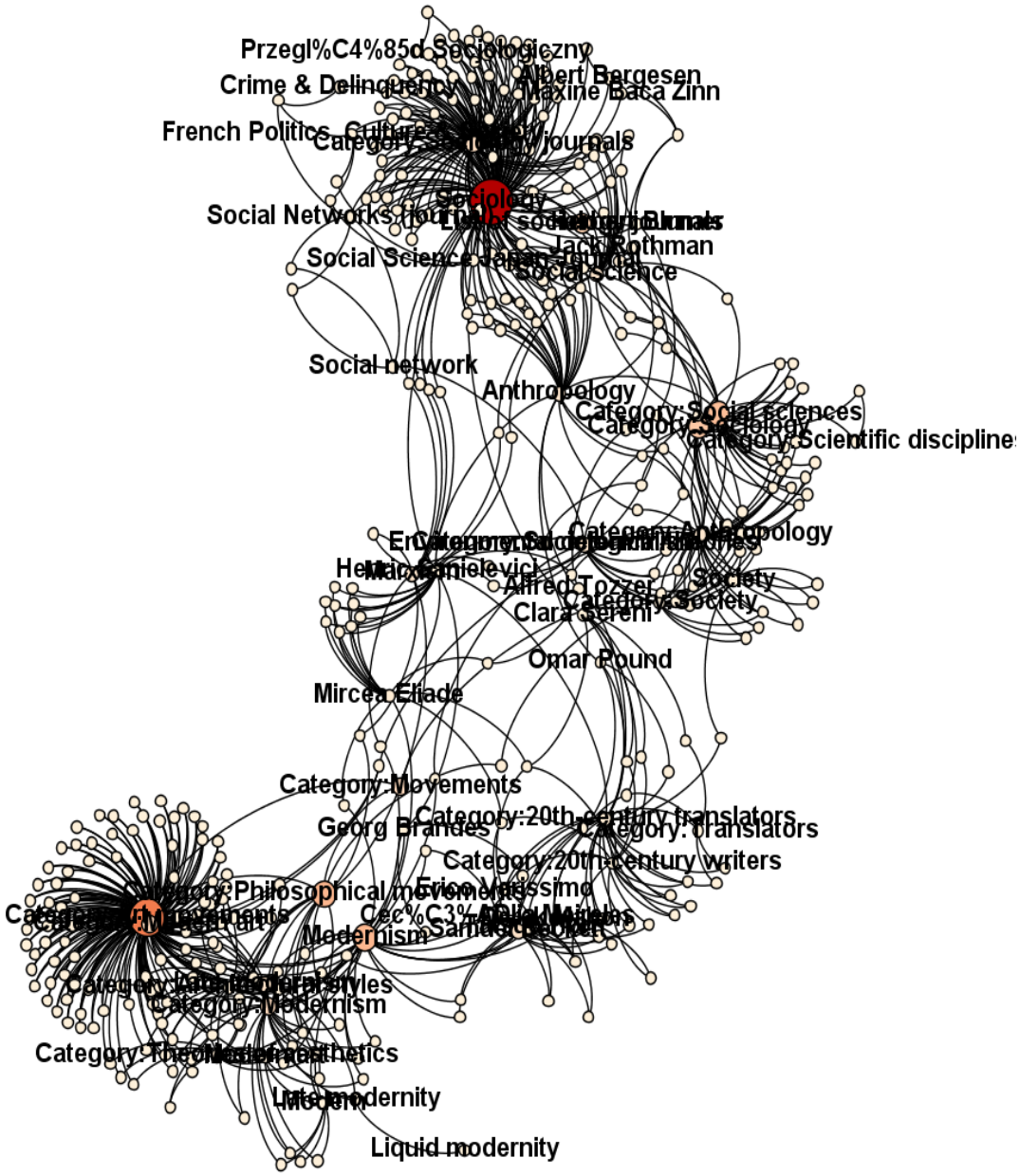
Fonte: Elaboração própria, por intermédio do protótipo e apoio do software Gephi

Figura 61 – Rede de informação referente à sexta iteração do algoritmo



Fonte: Elaboração própria, por intermédio do protótipo e apoio do software Gephi

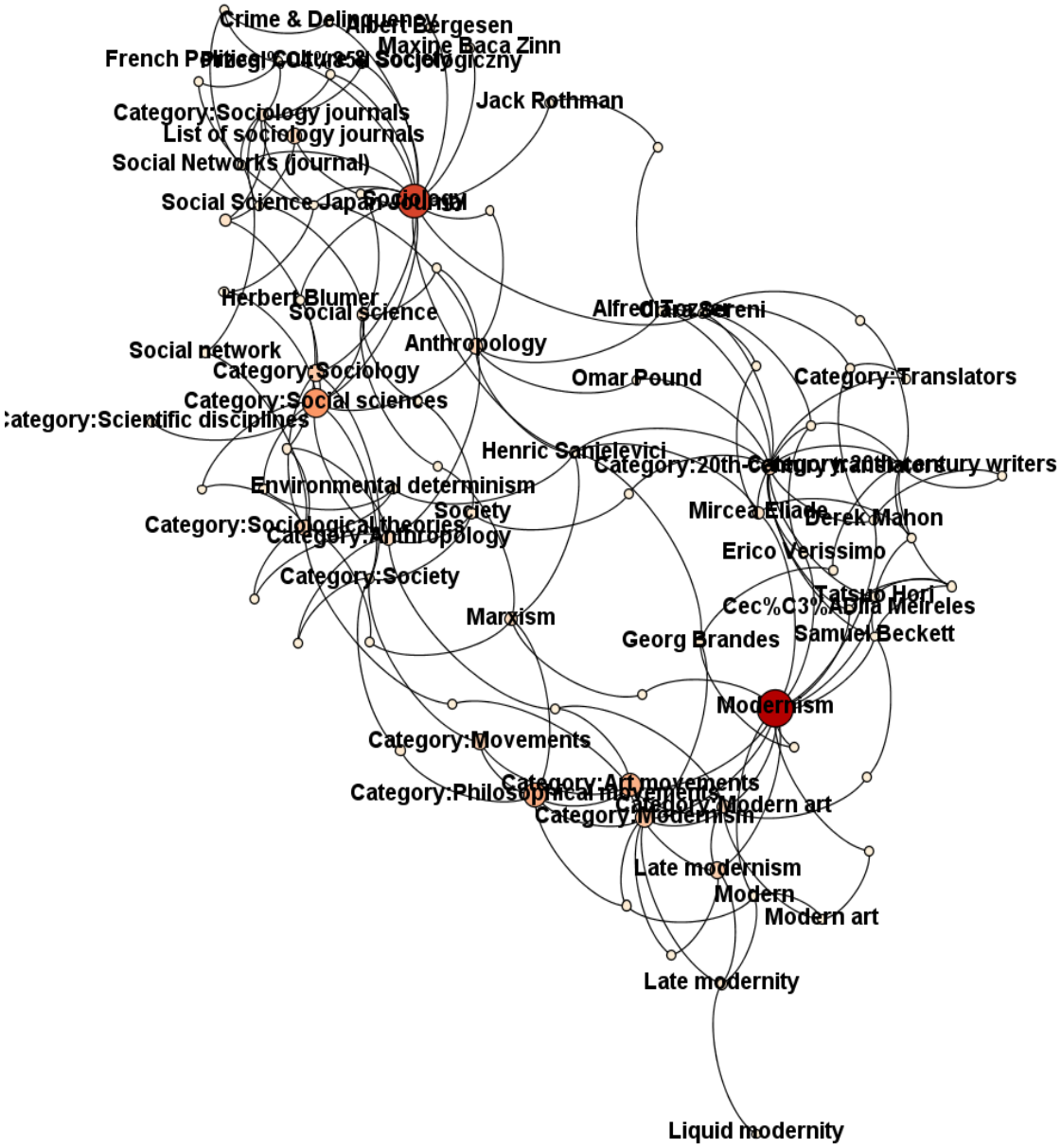
Figura 62 – Rede de informação referente à fase pós iterações e aplicação k-core



Fonte: Elaboração própria, por intermédio do protótipo e apoio do software Gephi

O primeiro estágio de redução da rede com aplicação de 2-core tem impacto grande sobre a rede, pois ela diminui de 6.023 nós, Figura 61, para 383 nós, Figura 63 e Quadro 8 em '*** Intermediate stage 1 (apply k-core) ***', mantendo os nós com grau 2 acima, que representam nós mais conectados do que aqueles eliminados com grau um.

Figura 63 – Rede de informação referente à fase de seleção dos nós principais

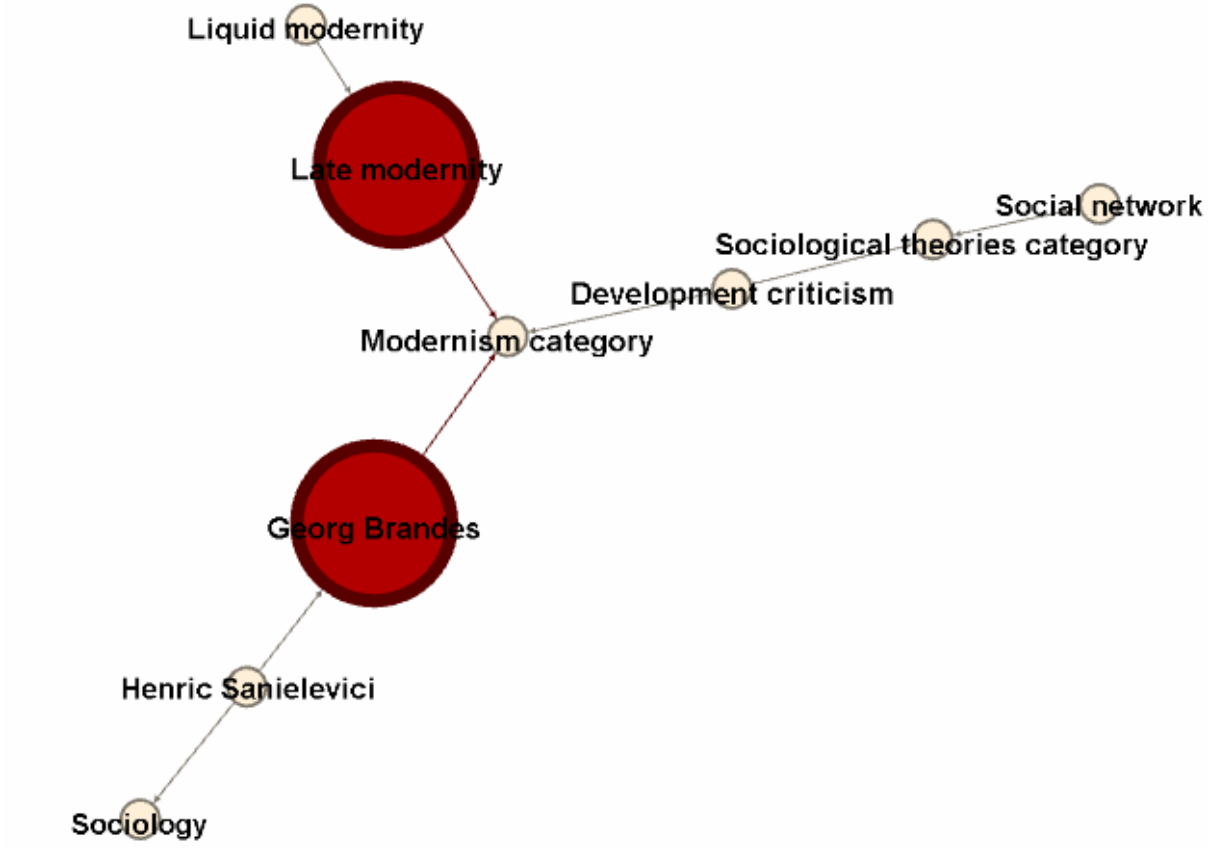


Fonte: Elaboração própria, por intermédio do protótipo e apoio do software Gephi

O próximo estágio de redução da rede consiste na reunião dos nós ranqueados e selecionados em cada iteração, 49 ao todo, para a formação de uma nova rede, sendo que os caminhos mínimos entre eles são conservados. Assim, a rede foi reduzida para 85 nós, como mostra a Figura 63 e o Quadro 8 em '*** Intermediate stage 2 (build head nodes list to calculate paths in the final stage)'. Dessa forma, foram conservados os nós mais importantes no contexto do fluxo informacional entre os termos de consulta do usuário e os selecionados

ao longo de todo o processo. Essa já é uma rede que representa um bom resultado com relações interessantes, contudo, por questões de legibilidade, ela ainda é grande para ser mapeada num mapa conceitual que, nesse caso, a meta é um máximo de 5 a 10 conceitos.

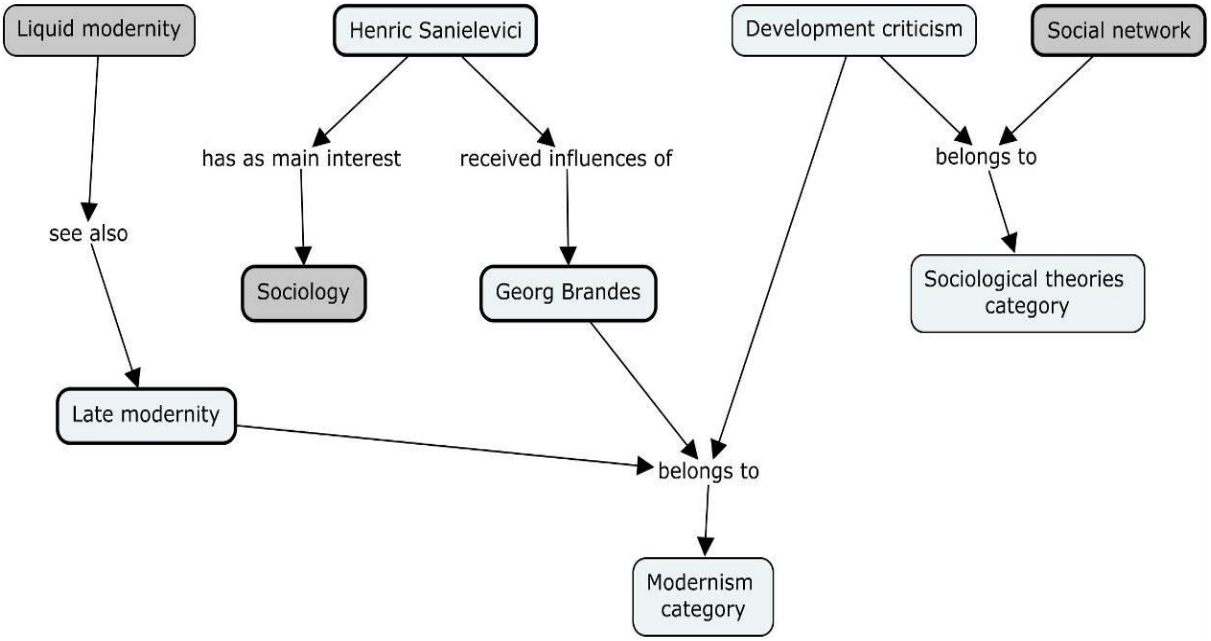
Figura 64 – Rede de informação referente ao último estágio do algoritmo



Fonte: Elaboração própria, por intermédio do protótipo e apoio do software Gephi

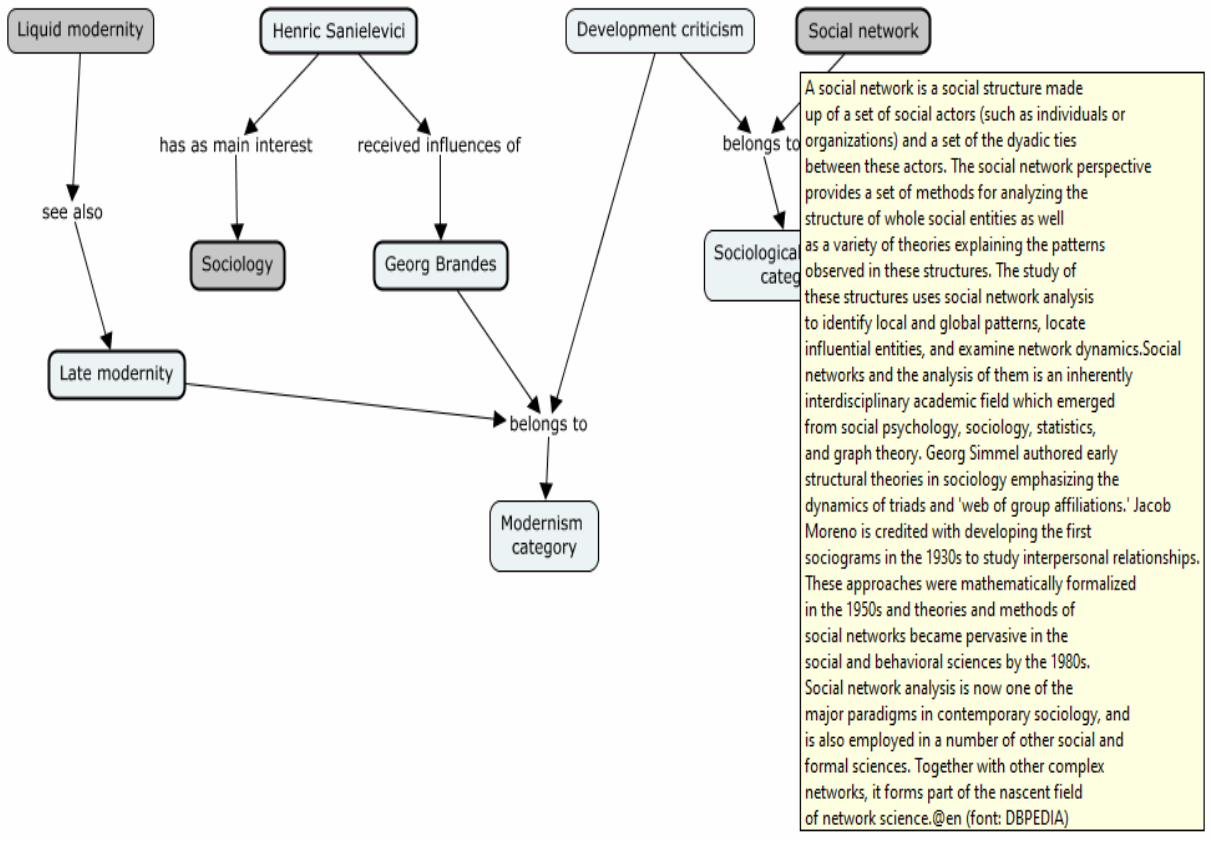
O último ciclo de redução da rede de informação acontece pela seleção e eliminação dos nós com maior valor de *eccentricity*, desde que os caminhos mínimos entre os termos de consulta do usuário sejam mantidos. Assim, a rede de informação fica com o tamanho de 9 nós, Figura 64, e, finalmente, ela é mapeada para o mapa conceitual resultante, Figura 65. Usando o software CmapTools, é possível visualizar o mapa, inclusive os *hints* dos conceitos destacados numa caixa de espessura maior, como é o caso do conceito ‘Social network’ na Figura 66. O *hint* é simplesmente coletado na DBpedia quando há predicado, na tripla RDF, indicando *abstract*.

Figura 65 – Mapa conceitual resultante para os termos ‘Sociology’, ‘Liquid modernity’ e ‘Social network’



Fonte: Elaboração própria, por intermédio do protótipo e apoio do software CmapTools

Figura 66 – Mapa conceitual resultante, com exemplo de *hint* para ‘Social network’



Fonte: Elaboração própria, por intermédio do protótipo e apoio do software CmapTools

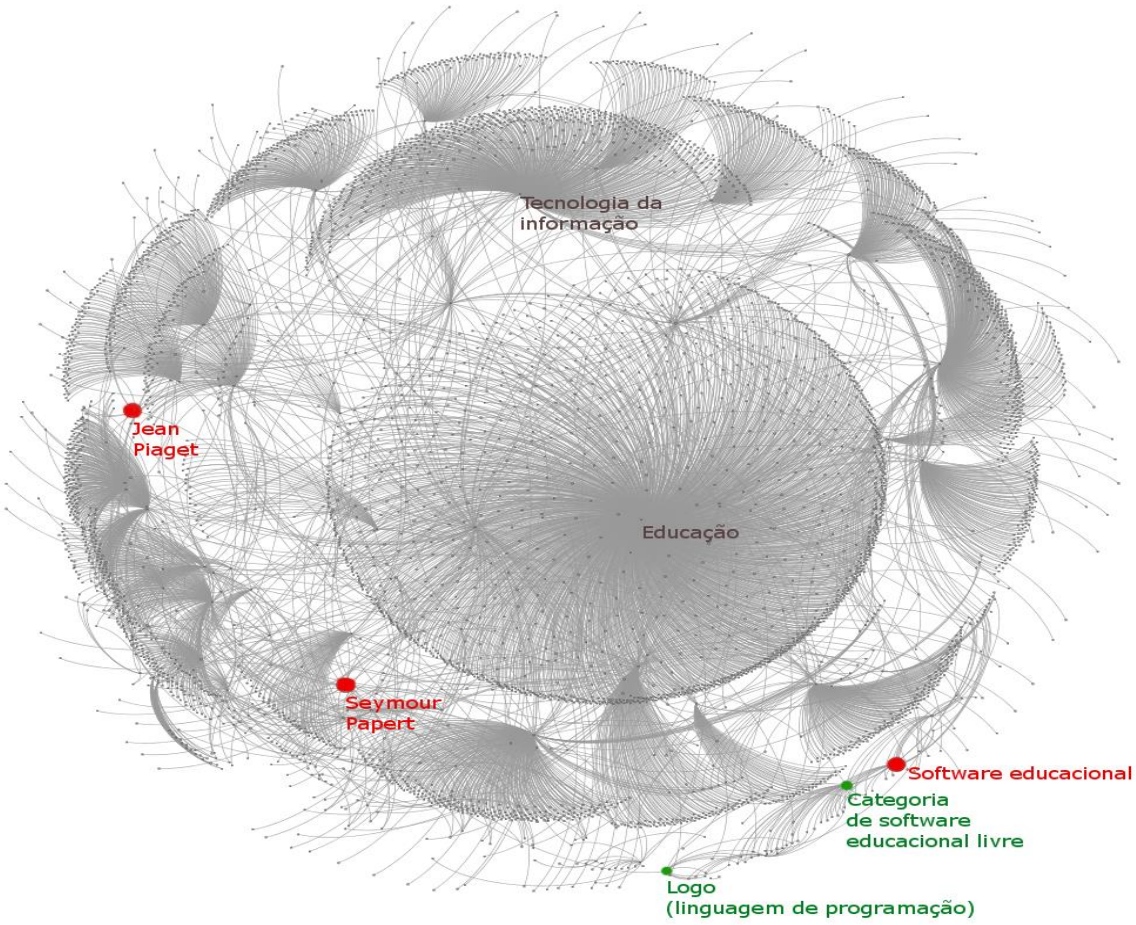
Quadro 8 – *Log* resumido da execução

```
*** Iteration 0 ***
- RDFs - 1026 new RDFs triples collected.
- Graph - 657 new nodes, 657 new edges.
- Connected components: 2.
- Level of relationship between original concepts: 33% complete.
- Selection of 10 new concepts to insert.
*** Iteration 1 ***
- RDFs - 408 new RDFs triples collected.
- Graph: 324 new nodes, 324 new edges.
- Total Graph: 981 nodes, 989 edges.
- Connected components: 2.
- Level of relationship between original concepts: 33% complete.
- Selection of 14 new concepts to insert.
*** Iteration 2 ***
- RDFs - 1025 new RDFs triples collected.
- Graph: 711 new nodes, 711 new edges.
- Total Graph: 1692 nodes, 1740 edges.
- Connected components: 1.
- Level of relationship between original concepts: 100% complete.
- Selection of 16 new concepts to insert.
*** Iteration 3 ***
- RDFs - 3451 new RDFs triples collected.
- Graph: 2757 new nodes, 2757 new edges.
- Total Graph: 4449 nodes, 4731 edges.
- Connected components: 1.
- Level of relationship between original concepts: 100% complete.
- Selection of 16 new concepts to insert.
*** Iteration 4 ***
- RDFs - 709 new RDFs triples collected.
- Graph: 501 new nodes, 501 new edges.
- Total Graph: 4950 nodes, 5275 edges.
- Connected components: 1.
- Level of relationship between original concepts: 100% complete.
- Selection of 6 new concepts to insert.
*** Iteration 5 ***
- RDFs - 1293 new RDFs triples collected.
- Graph: 1073 new nodes, 1073 new edges.
- Total Graph: 6023 nodes, 6413 edges.
- Connected components: 1.
- Level of relationship between original concepts: 100% complete.
- Selection of 6 new concepts to insert.
*** Intermediate stage 1 (apply k-core) ***
- 2-core filter algorithm  - 5640 deleted nodes (8 selected concepts) and 5638
deleted edges
- Remained Stream Graph: 383 nodes, 383 edges.
- Connected components: 1.
- Level of relationship between original concepts: 100% complete.
*** Intermediate stage 2 (build head nodes list to calculate paths in the final
stage) ***
- Head nodes from original concepts and selected concepts: 49 nodes.
- Filtering of graph: only nodes and edges belong to shortest path between head
nodes - 1176 paths found.
- Remained Stream Graph: 85 nodes, 163 edges.
- Resultant Graph: 85 nodes, 85 edges in the graph structure.
*** Intermediate stage 3 (remove nodes steadying unique connected component ***
- Removal of nodes, from graph with 85 nodes.
- ...
- Total concepts:
  6 remaining concepts + 3 original concepts = 9 total concepts  (goal 5 to 10)
  Connected component count: 1 (base: 1)
  Stream Graph: 9 nodes, 8 edges
*** Final stage (building concept map) ***
- Concept map with 8 proposition.
```

Fonte: Elaboração própria, por intermédio do protótipo

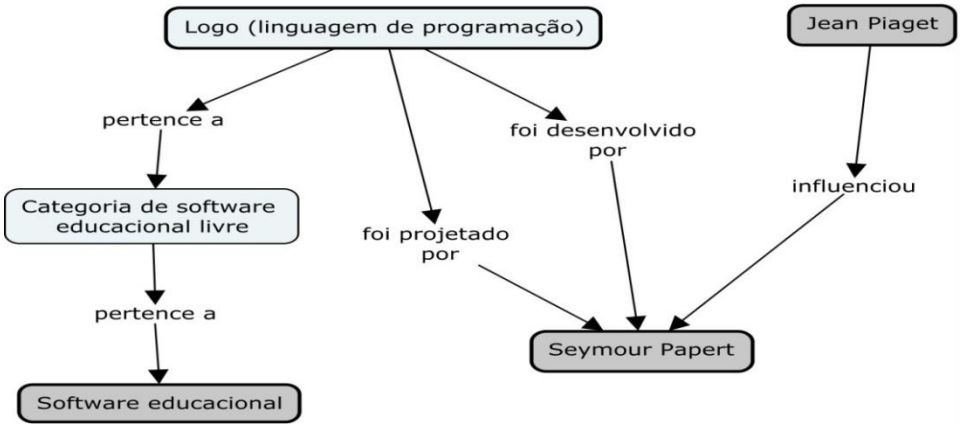
A Figura 67 mostra outro exemplo de rede expandida após oito ciclos de retroalimentação e com alguns milhares de nós e conexões, advindos de uma consulta sobre os termos ‘Jean Piaget’, ‘Software educacional’ e ‘Seymour Papert’. Essa rede, após redução, originou o mapa conceitual resultante mostrado na Figura 68, que foi um dos mapas avaliados na validação com usuários. Os outros conceitos do mapa, ‘Logo (linguagem de programação)’ e ‘Categoria de software educacional livre’, oferecem uma possibilidade de relacionamento entre os termos de consulta por intermédio de ranqueamentos e seleção no processamento do modelo sobre a rede expandida da Figura 67. Observa-se ainda na rede expandida, destaque para os nós ‘Educação’ e ‘Tecnologia da informação’, que, apesar de possuírem as duas maiores quantidades de conexões, não foram selecionados para o mapa resultante, pois o algoritmo do modelo tem critérios que vão além de uma simples escolha como essa, privilegiando, principalmente, o nível de intermediação entre eles.

Figura 67 – Rede de informação expandida com 4.285 nós e 4.909 conexões a partir dos termos de consulta ‘Jean Piaget’, ‘Software educacional’ e ‘Seymour Papert’



Fonte: Elaboração própria, por intermédio do protótipo e apoio do software Gephi

Figura 68 – Mapa conceitual resultante dos termos ‘Jean Piaget’, ‘Software educacional’ e ‘Seymour Papert’



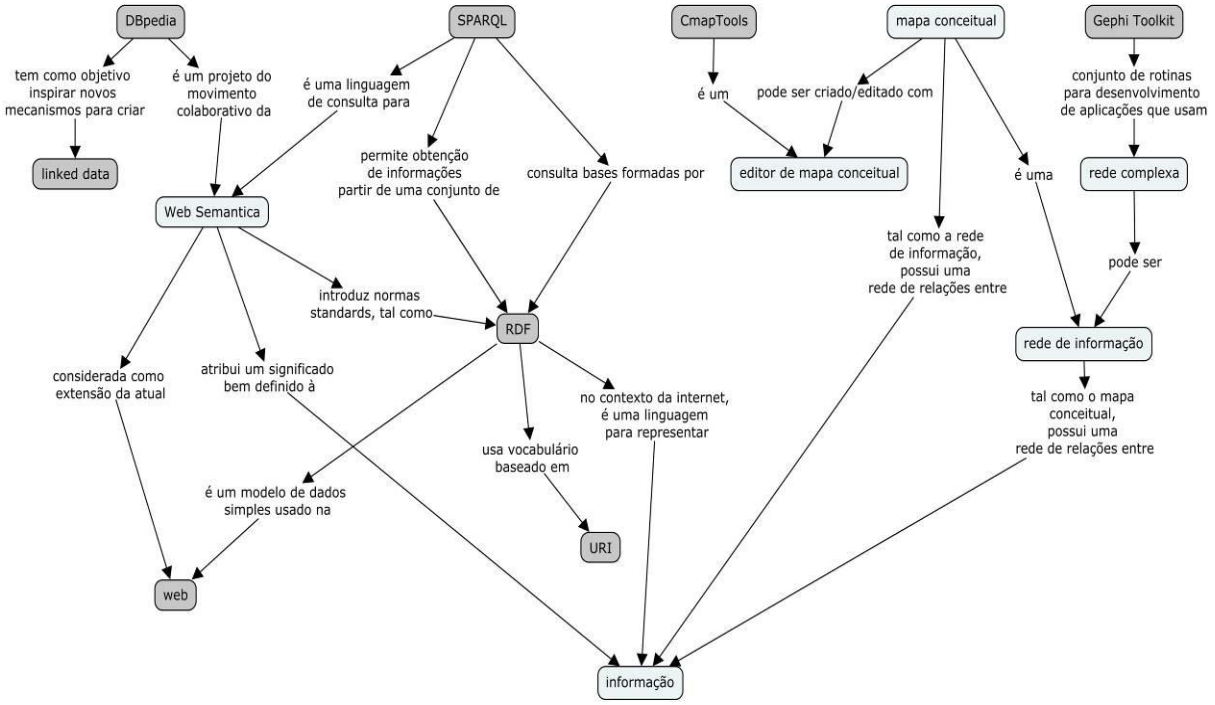
Fonte: Elaboração própria, por intermédio do protótipo e apoio do software CmapTools

4.5.6 Testes piloto em uma base de conhecimento privada

Com o intuito de diversificar os testes para uma base diferente da DBpedia, foram elaborados testes piloto com o protótipo do modelo aprimorado sobre uma base de conhecimento privada, contendo uma rede de informação formada por 850 conceitos e 1567 conexões. Essa base contém informações, principalmente, sobre as áreas: Ciência da Informação, Web Semântica, Dados Abertos Ligados, Ciência das Redes, Mapas Conceituais e Recuperação de Informação. Ela não é pública e não tem predicados formados a partir de ontologias, tal com a DBpedia, mas, os seus predicados são mais descritivos, contendo explicações mais didáticas sobre relacionamentos descritos por suas respectivas proposições, isto é, pelo relacionamento entre os seus respectivos pares de conceitos.

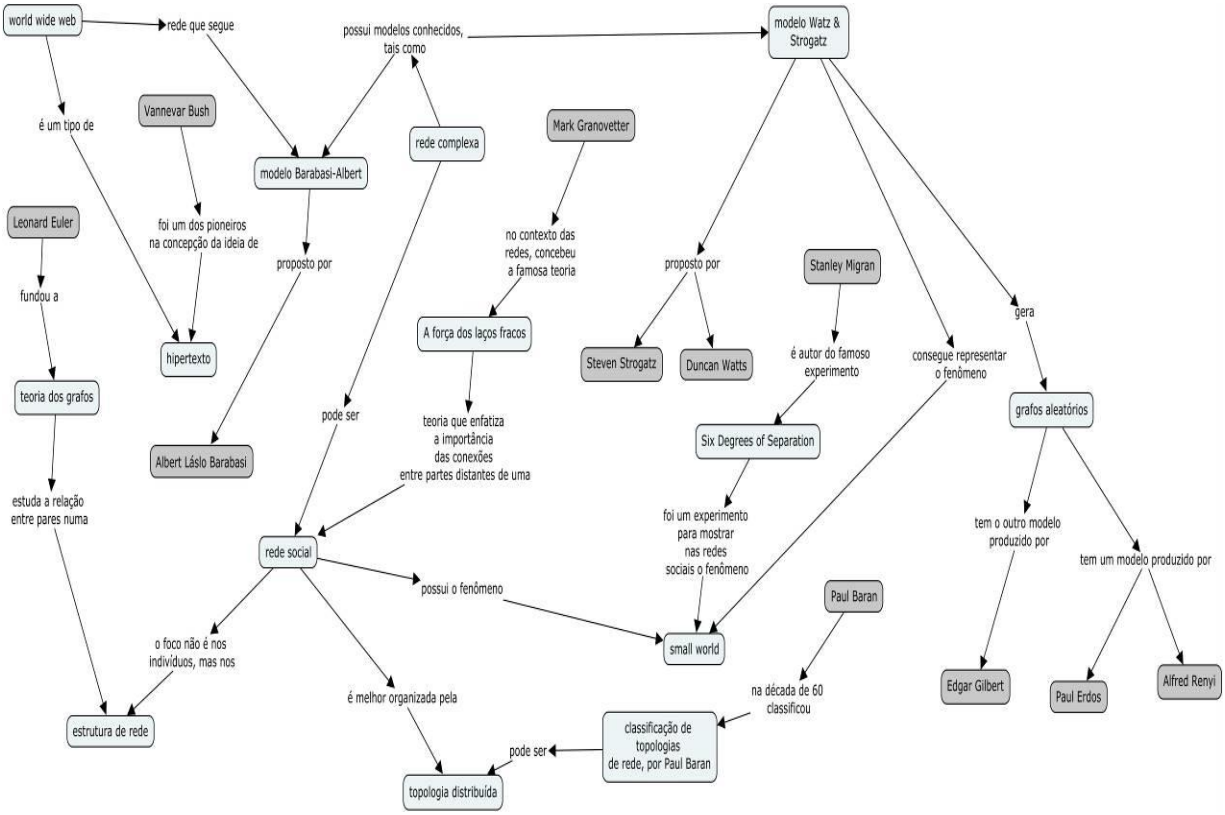
A Figura 69 e a Figura 73 mostram 5 mapas conceituais resultantes do processamento feito a partir de 5 conjuntos arbitrários de termos escolhidos dentre os 850 disponíveis. A escolha dos termos tentou, em alguns casos, explorar a diversidade das áreas de conhecimento para verificar o quanto o modelo consegue estabelecer relacionamentos entre os termos, mesmo que conceitualmente distantes, e também distantes no contexto de conexões da rede de informação. Os conceitos em caixas de cor de fundo cinza são os termos de consulta do usuário.

Figura 69 – Mapa conceitual resultante do primeiro teste na base de conhecimento privados



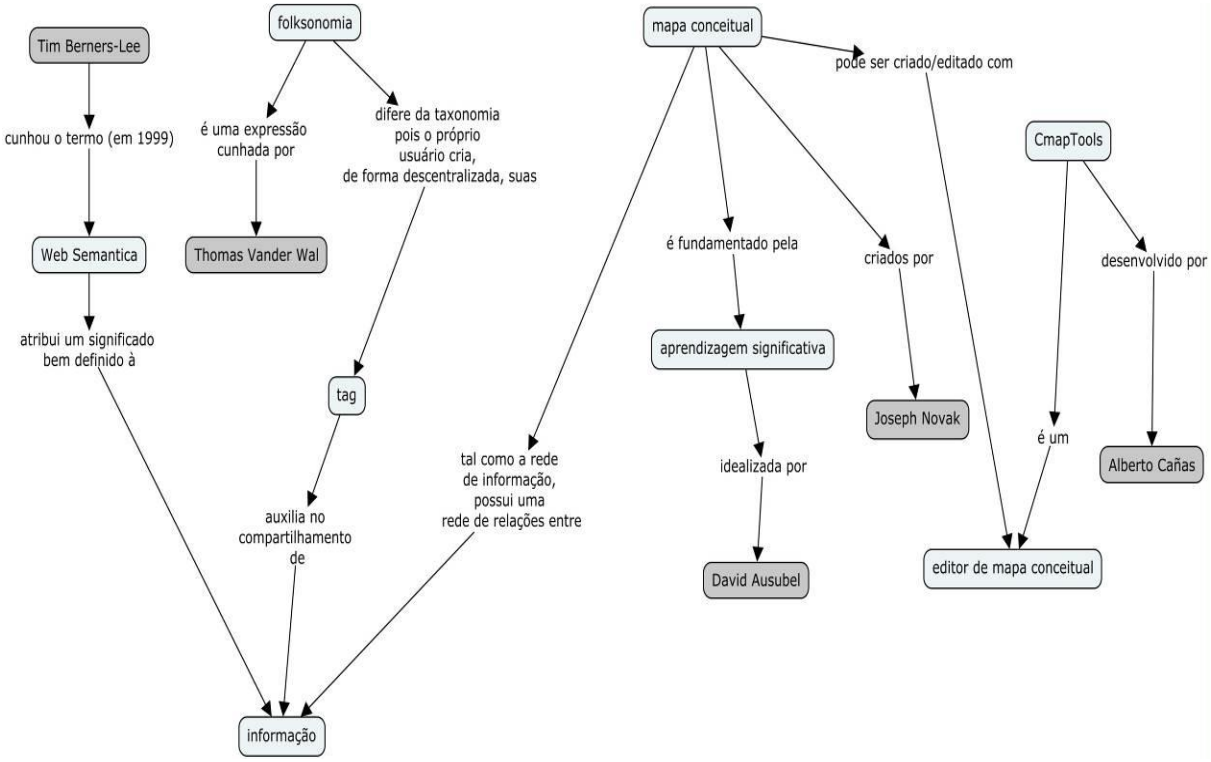
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 70 – Mapa conceitual resultante do segundo teste na base de conhecimento privados



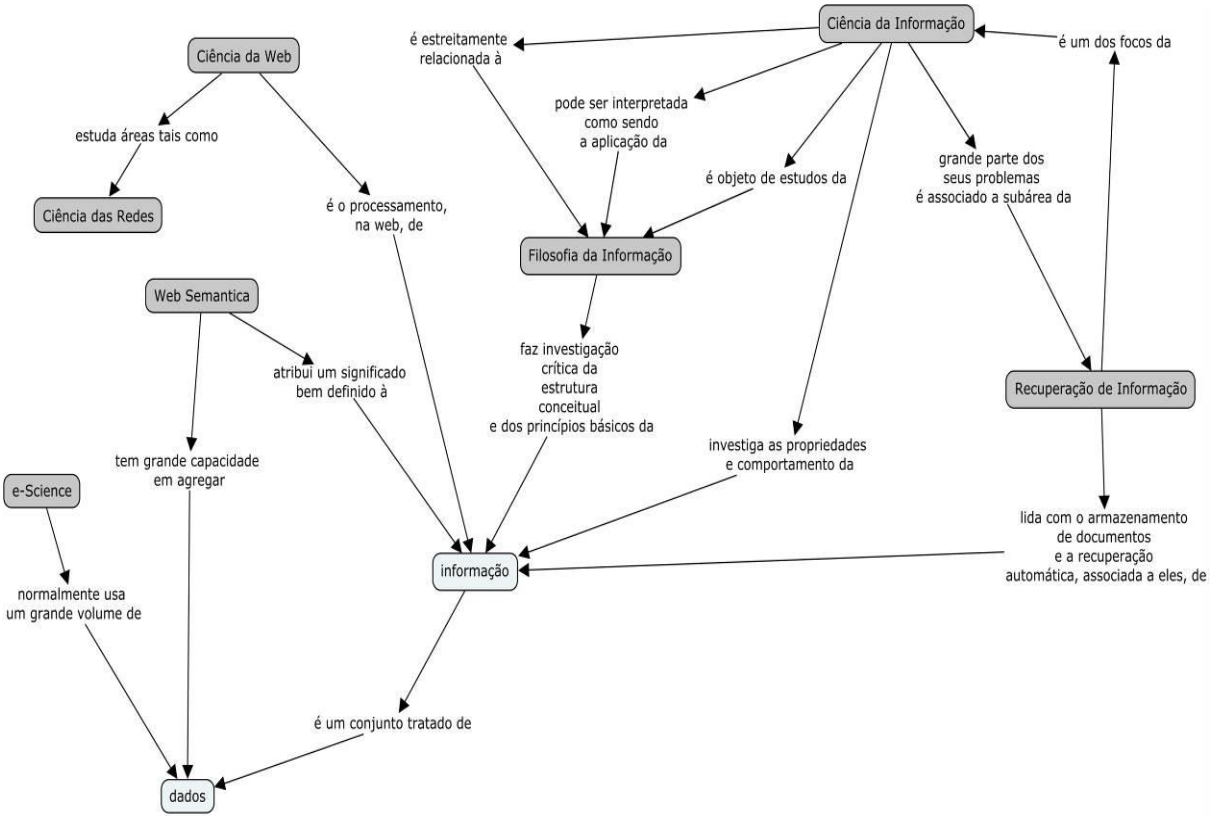
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 71 – Mapa conceitual resultante do terceiro teste na base de conhecimento privados



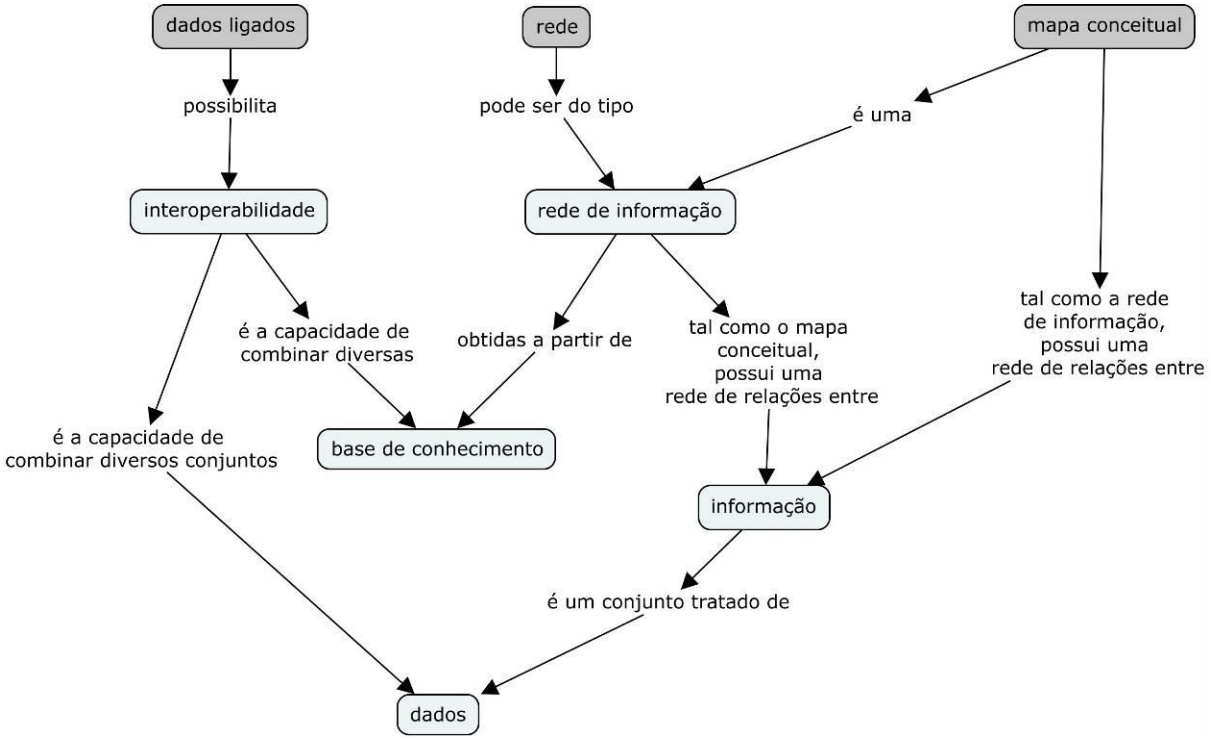
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 72 – Mapa conceitual resultante do quarto teste na base de conhecimento privados



Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 73 – Mapa conceitual resultante do quinto teste na base de conhecimento privados



Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

O mapa conceitual resultante do primeiro teste, Figura 69, é proveniente dos termos ‘DBpedia’, ‘SPARQL’, ‘linked data’, ‘CmapTools’, ‘Gephi Toolkit’, ‘RDF’, ‘web’ e ‘URI’. Houve casos de descoberta de relacionamentos entre termos distantes, isto é, de áreas diferentes. Por exemplo, ‘CmapTools’ e ‘linked data’ precisou de 5 conceitos intermediários. Por outro lado, conceitos como ‘web’ e ‘RDF’ tiveram ligação direta por estarem conceitualmente próximos.

O mapa conceitual do segundo teste, Figura 70, possui 25 conceitos finais e conseguiu estabelecer conexões com 11 personalidades da área Ciência das Redes: ‘Leonard Euler’, ‘Vannemar Bush’, ‘Albert Láslo Barabasi’, ‘Mark Granovetter’, ‘Steven Strogatz’, ‘Ducan Watts’, ‘Stanley Migran’, ‘Paul Baran’, ‘Edgar Gilbert’, ‘Paul Erdos’ e ‘Alfred Renyi’.

O mapa conceitual do terceiro teste, Figura 71, relaciona autores de duas áreas: web (‘Tim Berners-Lee’ e ‘Thomas Vander Wal’) e mapas conceituais (‘Joseph Novak’, ‘David Ausubel’ e ‘Alberto Cañas’), com 13 conceitos finais.

A quantidade de conceitos intermediários pode ser minimizada nas configurações do algoritmo, como por exemplo, o mapa conceitual do quarto teste, Figura 72, possui apenas 2 conceitos intermediários com 7 termos de consulta do usuário: ‘Ciência das Redes’, ‘Ciência

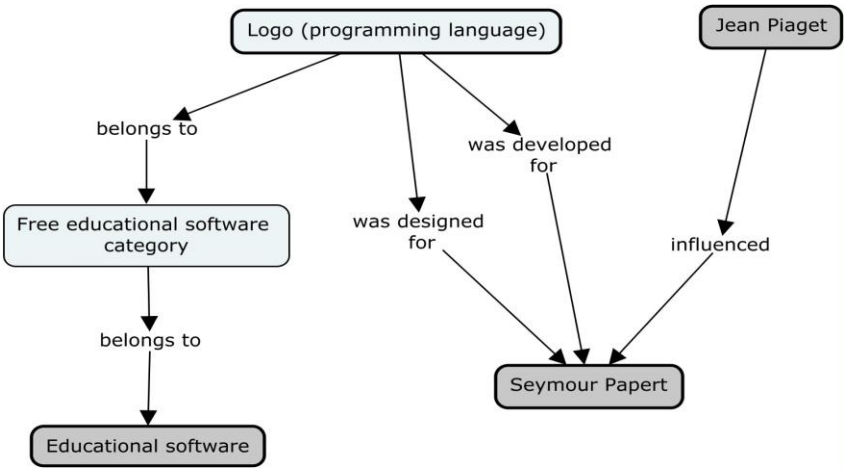
da Web', 'Web Semântica', 'e-Science', 'Filosofia da Informação', 'Ciência da Informação', Recuperação de Informação'.

O mapa conceitual do quinto teste, Figura 73, por outro lado teve a quantidade de conceitos intermediários configurada para descobrir 5 conceitos intermediários para apenas 3 termos de consulta do usuário: 'dados ligados', 'rede' e 'mapa conceitual'. Dessa forma, aparecem relacionamentos que não são necessários para interligar todos os termos num único componente conectado, tal como 'interoperabilidade é a capacidade de interligar diversos conjuntos de dados'. Porém, são relações que podem enriquecer o conhecimento do usuário ao ler o mapa resultante.

4.5.7 Testes piloto numa base de dados abertos ligados: DBpedia

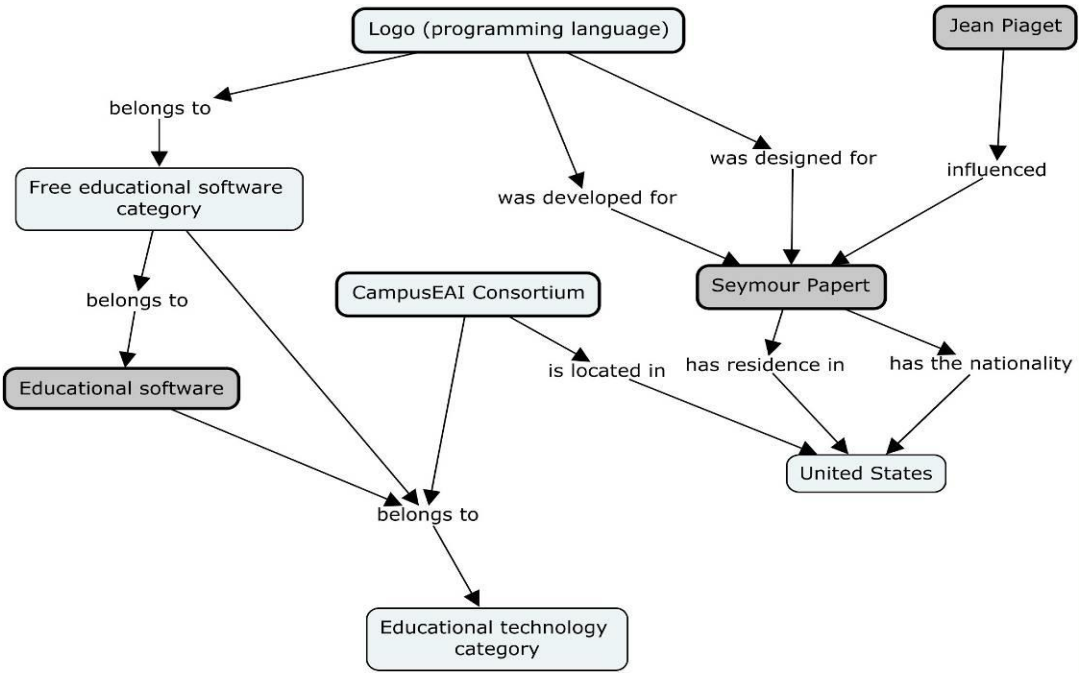
A DBpedia é uma base de dados abertos ligados, tal como já caracterizada na subseção 2.3.5 do referencial teórico e na seção 3.5.1 da metodologia. Uma vez que a DBpedia seria usada na validação do modelo com os usuários, foram elaborados vários testes piloto com o protótipo. Quatro deles são apresentados aqui nessa subseção. A Figura 74 até a Figura 77 mostram os mapas resultantes desses testes sobre conjuntos arbitrários de termos.

Figura 74 – Mapa conceitual resultante dos termos 'Jean Piaget', 'Educacional software' e 'Seymour Papert' com dois conceitos intermediários



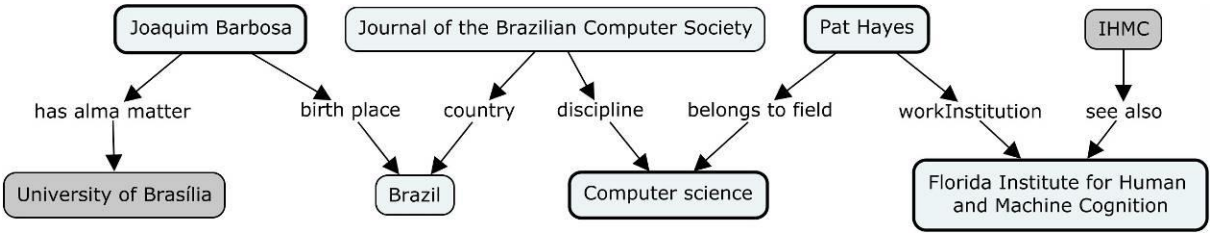
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 75 – Mapa conceitual resultante dos termos ‘Jean Piaget’, ‘Educacional software’ e ‘Seymour Papert’ com cinco conceitos intermediários



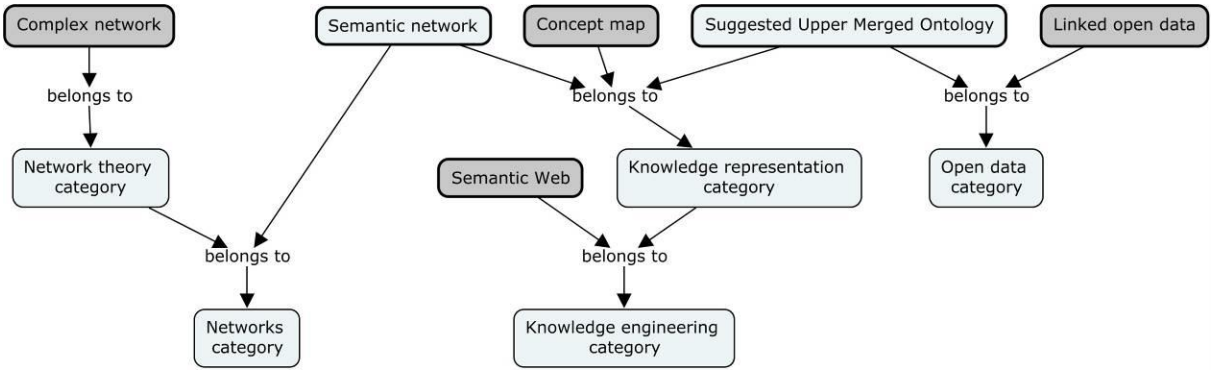
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 76 – Mapa conceitual resultante dos termos ‘University of Brasilia’ e ‘IHMC’



Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 77 – Mapa conceitual resultante dos termos ‘Complex network’, ‘Concept map’, ‘Semantic Web’ e ‘Linked open data’



Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

O teste feito sobre o conjunto de termos ‘Jean Piaget’, ‘Educacional software’ e ‘Seymour Papert’ gerou dois mapas, um configurado para revelar 2 conceitos novos no mapa conceitual final, Figura 74, e outro com uma quantidade maior de conceitos novos, Figura 75. Esse ajute é realizado por uma configuração opcional que é feita na execução do protótipo.

O teste com os termos ‘University of Brasilia’ e ‘IHMC’, Figura 76, mostra a dependência do funcionamento do modelo com a qualidade da base de conhecimento, pois, apesar de existirem relações diretas entre as duas instituições, como por exemplo, as duas trabalham com pesquisa em Ciência da Computação, isso não apareceu no mapa. É óbvio que, primeiramente, alguém deveria alimentar a base de conhecimento com essa relação. Assim, o sistema achou um caminho entre os dois termos, um tanto quanto inusitado, mas dentro do escopo existente na DBpedia. É também importante destacar que o termo ‘University of Brasilia’ se conectou com o conceito ‘Brazil’ por intermédio de ‘Joaquim Barboza’. Porém, poder-se-ia questionar, porque não outra pessoa? A resposta para essa questão e outras da mesma natureza é que a escolha sempre se dá sobre aquele nó com mais importância na rede, sendo que o critério de medição da importância é determinado pelos algoritmos de ranqueamento pertencentes ao algoritmo do modelo.

A Figura 77 mostra um mapa resultante dos termos ‘Complex network’, ‘Concept map’, ‘Semantic Web’ e ‘Linked open data’. Esse teste, apesar de ter conseguido relacionar os 4 termos de consulta do usuário, ele demonstra excesso de simplicidade nas proposições advindas da base de dados, formadas somente pela frase de ligação ‘belongs to’ (pertence a). Por outro lado, esse tipo de relação revela um mapa conceitual hierárquico no contexto das áreas do conhecimento, onde os dois conceitos troncos são ‘Networks category’ (categoria de redes) e ‘Knowledge engineering category’ (categoria de engenharia do conhecimento).

4.6 Resultados da validação do modelo aprimorado

Essa seção apresenta os resultados da validação do modelo aprimorado com um grupo de usuários, por intermédio do protótipo implementado. São apresentados os termos de consulta coletados, os resultados das avaliações realizadas diretamente pelos usuários e os resultados das avaliações realizadas com o método para medição da qualidade da informação recuperada (apresentado na subseção 3.5.5).

4.6.1 Termos de consulta coletados

O Quadro 9 e o Quadro 10 mostram os conjuntos de termos fornecidos pelos usuários, chamados também de termos base. Não é levada em consideração a ordem dos termos ou qualquer tipo de peso entre eles. Foram ao todo 32 conjuntos de termos base. O método usado nessa coleta de dados foi apresentado na subseção 3.5.4. Os mapas conceituais resultantes de cada consulta estão disponíveis no APÊNDICE H.

Quadro 9 – Conjuntos de três termos fornecidos pelos usuários

Ident.	1º termo	2º termo	3º termo
01	Developmental psychology	Chance (philosophy)	Meaning (psychology)
02	Idiosincrasia	Metacognição	Educação
04	Cidadania	Inclusão	Diversidade
05	Notação musical	Teoria musical	Música
06	Ambiente virtual de aprendizagem	Tecnologia educacional	Robótica educacional
07	Mundo líquido	Sociologia	Rede social
09	Processo de Desenvolvimento de Software	Software Educacional	Objeto de Aprendizagem
10	Desejo	Arthur Schopenhauer	Vegetarianismo
12	Joan Miró	Surrealismo	Artista
13	Knowledge organization	Concept map	Information science
14	Data mining	Artificial intelligence	Machine learning
15	Baleia	Tubarão	Golfinho
16	Objeto de Aprendizagem	Tecnologia Educacional	Ciberespaço
17	Linguagem	Pesquisa	Conhecimento
18	Aprendizagem colaborativa	Área de Estudo	Avaliação
19	Educação	Aprendizagem	Avaliação

Fonte: Elaboração própria

Quadro 10 – Conjuntos de seis termos fornecidos pelos usuários

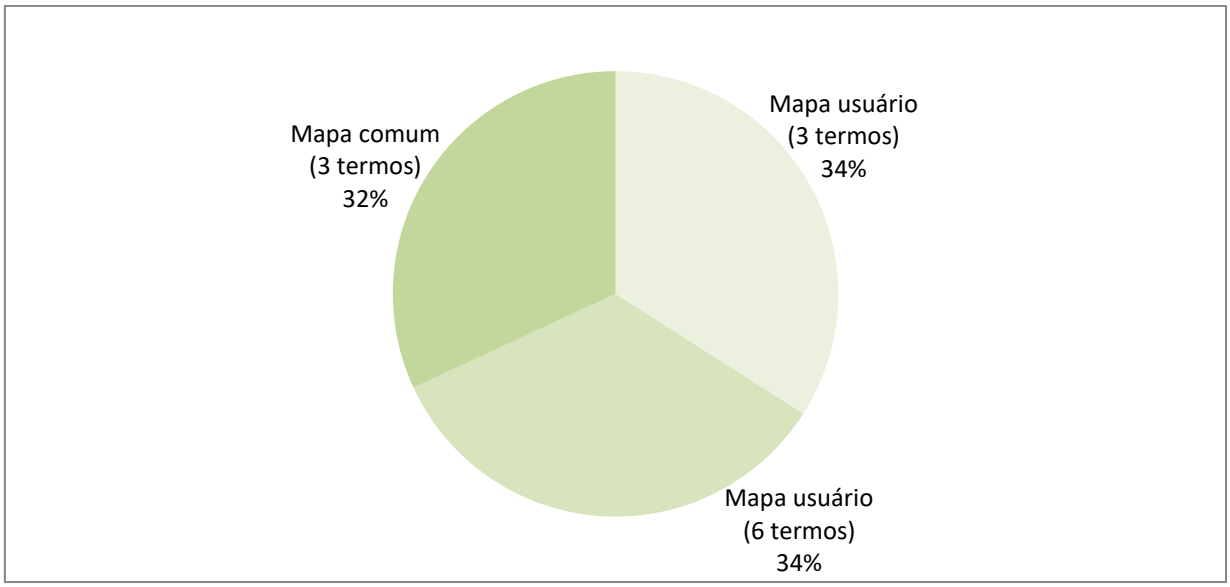
Id	1º termo	2º termo	3º termo	4º termo	5º termo	6º termo
01	Developmental psychology	Chance (philosophy)	Meaning (psychology)	Adult	Jean Piaget	Combinatorics
02	Mapa mental	Tony Buzan	Aprendizagem	Imagem	FreeMind	Memorização
04	Educação	Família	Sociedade	Responsabilidade social	Declaração Universal dos Direitos Humanos	Parceria
05	Educação	Currículo	Aprendizagem	Metodologia	Recurso (computação)	Avaliação
06	Artes marciais	Aikido	Ninjutsu	Habilidades de sobrevivência	Hapkido	Bushido
07	Gamification na educação	Jogo	Digital media	Formação docente	Teorias de Ensino	Autonomia
08	Diversidade (politica)	Respeito	Cultura	Currículo	Africano	Negro
10	Foucault	Poder	Biopolítica	Anátomo-política	Saúde	Controle

12	Educação infantil	Didática	Plano de Ensino	Pedagogia	Curriculum	Jean Piaget
13	Institutional repository	Digital preservation	Open access	Digital information	Institutional memory	Information
14	Data mining	Business intelligence	Analytic	Data warehouse	Knowledge discovery in databases	Online analytical processing
15	Degradação Ambiental	Área de Proteção	Reserva Natural	Sítio Arqueológico	Falésia	Dunas
16	Cavaleiros do Zoodíaco	MasterChef	Séries de TV	Futebol	Copa do Mundo FIFA	Olimpíadas 2016
17	Howard Gardner	Teoria das Inteligências Múltiplas	Aprendizagem	Habilidades	Perspective (cognitive)	Avaliação
18	Educação a Distância	Disciplina Escolar	Feedback	Interação	Professor	Estudante
19	Educação a Distância	Internet	Rede Social	Interação Social	Ambiente Virtual de Aprendizagem	Cooperação

Fonte: Elaboração própria

O Gráfico 7 apresenta a quantidade total de 47 avaliações realizadas, onde ‘mapa comum’ refere-se ao mapa conceitual avaliado por todos os usuários, apresentado no capítulo da metodologia, subseção 3.5.3 - Figura 37, e ‘mapa usuário’ refere-se aos mapas provenientes dos dois conjuntos de termos fornecidos pelos usuários, com três e seis termos.

Gráfico 7 – Quantidade de avaliações realizadas (total = 47)

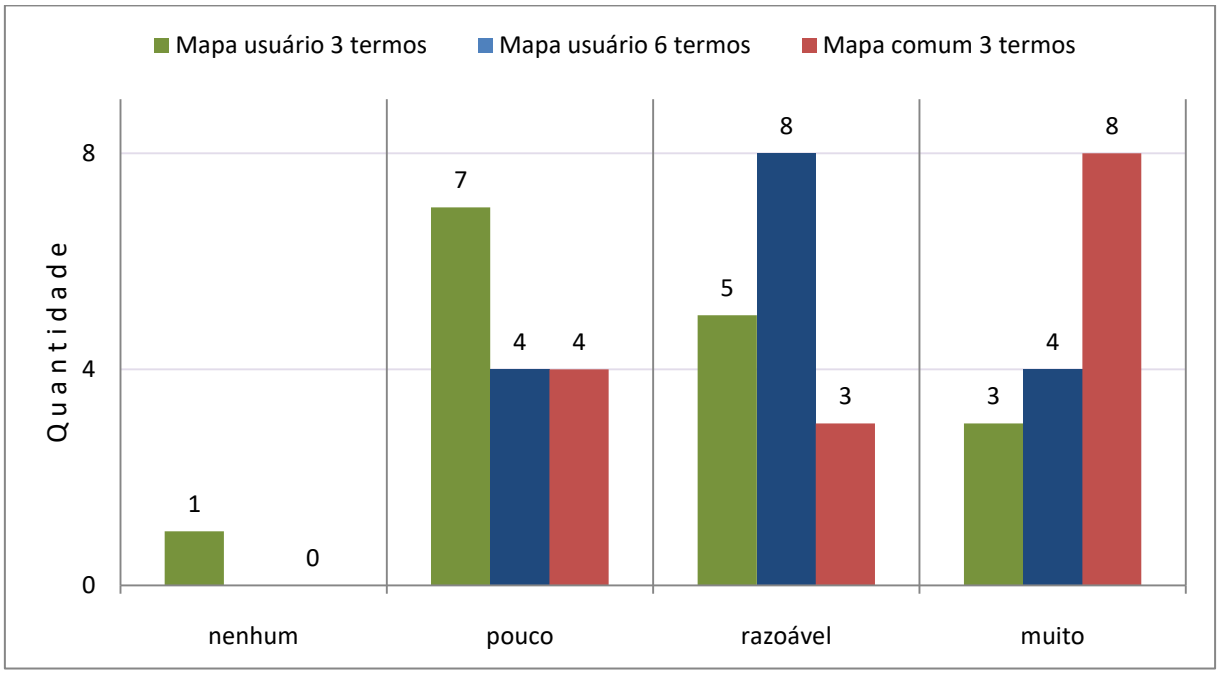


Fonte: Elaboração própria

4.6.2 Avaliações diretas dos usuários

Essa seção trata das avaliações feitas diretamente sobre os três mapas conceituais resultantes, segundo os seguintes três aspectos: **(i) o quanto o mapa auxilia o entendimento das relações entre os termos base**, **(ii) o quanto o mapa auxilia como ponto de partida para uma pesquisa sobre as relações entre os termos base**, **(iii) o quanto o mapa auxilia como ponto de partida na construção de um mapa conceitual sobre as relações entre os termos base**. Foram considerados, de acordo com o questionário de coleta das avaliações (APÊNDICE G), quatro níveis de resposta dos usuários: (1) **nenhum**, (2) **pouco**, (3) **razoável** e (4) **muito**. Dessa forma, o Gráfico 8, Gráfico 9 e Gráfico 10 contém um levantamento para cada um dos três aspectos avaliados, considerando de forma separada: '**mapa usuário 3 termos**' é o mapa conceitual resultante do grupo de três termos fornecidos pelo usuário, '**mapa usuário 6 termos**' é o mapa conceitual resultante do grupo de seis termos fornecidos pelo usuário, e '**mapa comum 3 termos**' é o mapa conceitual avaliado por todos os usuários (Figura 37) resultante de um conjunto arbitrário de três termos.

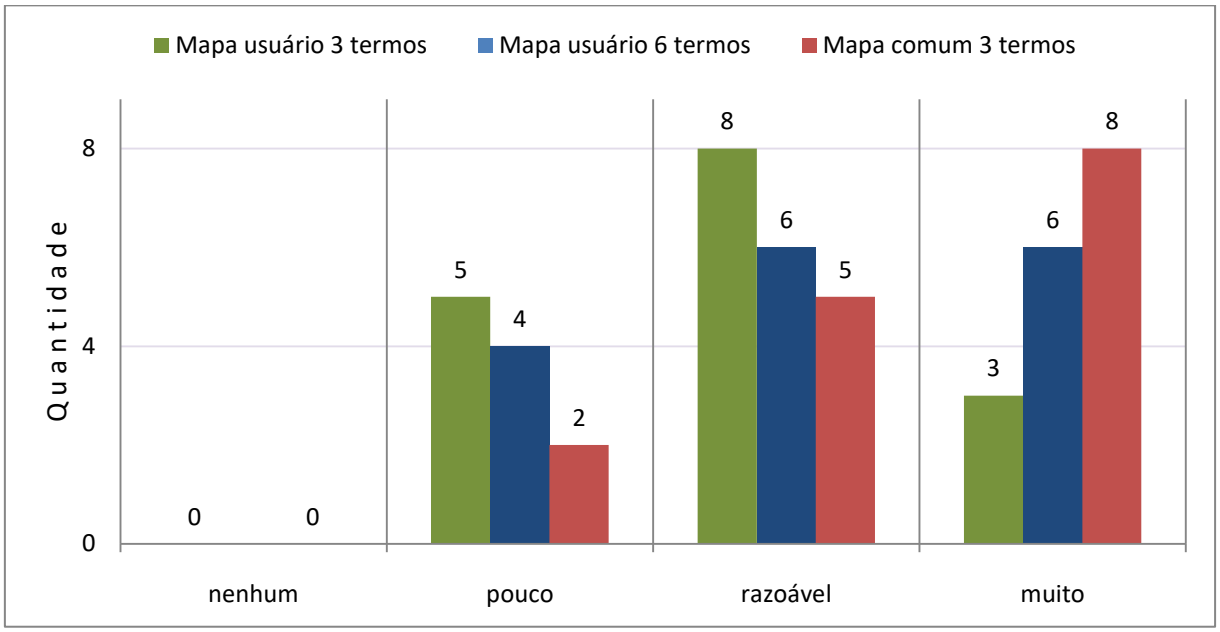
Gráfico 8 – Levantamento sobre o quanto o mapa conceitual auxilia o entendimento das relações entre os termos base



Fonte: Elaboração própria

Analisando o Gráfico 8, que faz o levantamento referente a quanto o mapa auxilia o entendimento das relações entre os termos base, observa-se que o mapa do usuário de 6 termos e o mapa comum de 3 termos foram melhores avaliados, pois a soma entre os níveis muito e razoável foi igual a 12 (4+8) indicações e 11 (8+3) indicações respectivamente, enquanto o mapa do usuário de três termos obteve 8 (3+5) indicações nesses mesmos níveis. De uma maneira geral, nos três mapas avaliados, a soma dos níveis muito e razoável obteve 31 (3+4+8+5+8+3) indicações contra 16 (7+4+4+1) indicações dos níveis pouco e nenhum, isto é, esse resultado sugere que os mapas avaliados proporcionaram um bom nível de auxílio para entendimento das relações entre os termos base.

Gráfico 9 – Levantamento sobre o quanto o mapa conceitual auxilia como ponto de partida para uma pesquisa sobre as relações entre os termos base

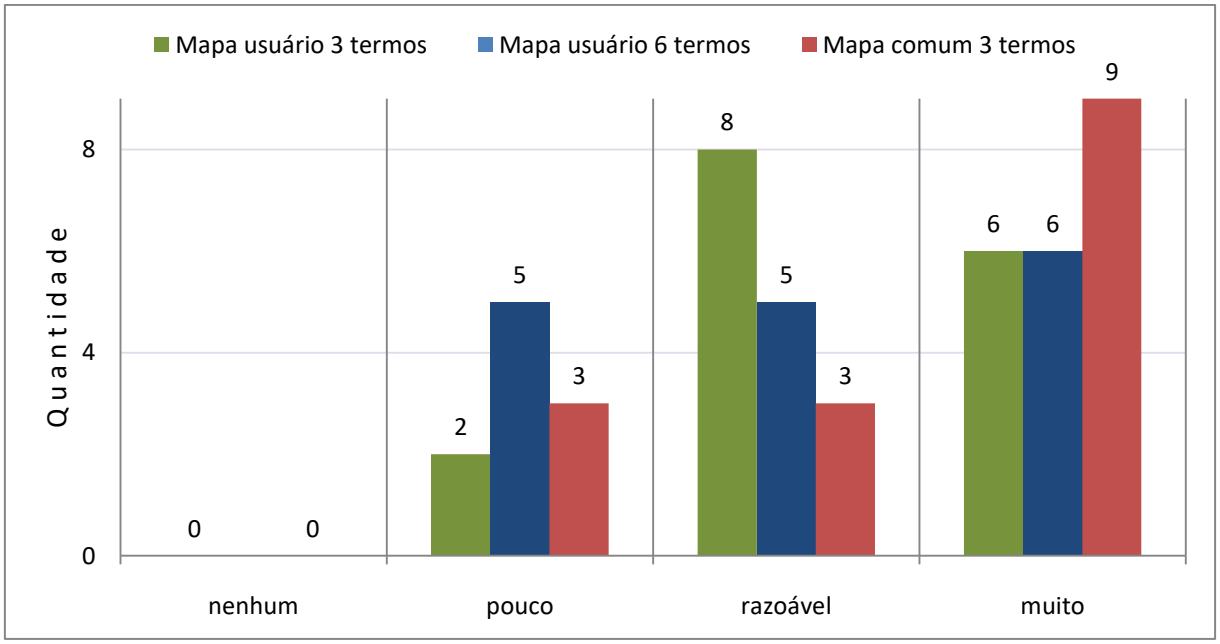


Fonte: Elaboração própria

A análise do Gráfico 9, que faz o levantamento referente a quanto o mapa conceitual serve como ponto de partida para uma pesquisa sobre as relações entre os termos base, apresenta uma similaridade entre os três mapas, pois a soma entre os níveis muito e razoável ficou muito próxima em cada uma deles, com 12 (8+5) indicações no mapa comum de 3 termos, com 12 (6+6) indicações no mapa do usuário com 6 termos, e com 11 (8+3) indicações no mapa do usuário com 3 termos. De uma maneira geral, nos três mapas avaliados, a soma dos níveis muito e razoável obteve 36 (3+6+8+8+6+5) indicações contra 11 (5+4+2) indicações dos níveis pouco e nenhum, isto é, esse resultado sugere que os mapas

avaliados proporcionaram um bom nível de auxílio como ponto de partida para uma pesquisa sobre as relações entre os termos base.

Gráfico 10 – Levantamento sobre o quanto o mapa conceitual auxilia como ponto de partida na construção de um mapa conceitual sobre as relações entre os termos base



Fonte: Elaboração própria

Na analise do Gráfico 10, que faz o levantamento sobre o quanto o mapa conceitual auxilia como ponto de partida na construção de um mapa conceitual sobre as relações entre os termos base, observa-se também, como no Gráfico 9, uma similaridade entre os três mapas, pois a soma entre os níveis muito e razoável ficou muito próxima em cada uma deles, com 12 (9+3) indicações no mapa comum de 3 termos, com 11 (6+5) indicações no mapa do usuário com 6 termos, e com 14 (6+8) indicações no mapa do usuário com 3 termos. De uma maneira geral, nos três mapas avaliados, a soma dos níveis muito e razoável obteve 37 (6+6+9+8+5+3) indicações contra 10 (2+5+3) indicações dos níveis pouco e nenhum, isto é, esse resultado sugere que os mapas avaliados proporcionam um bom nível de auxílio como ponto de partida na construção de um mapa conceitual sobre as relações entre os termos base.

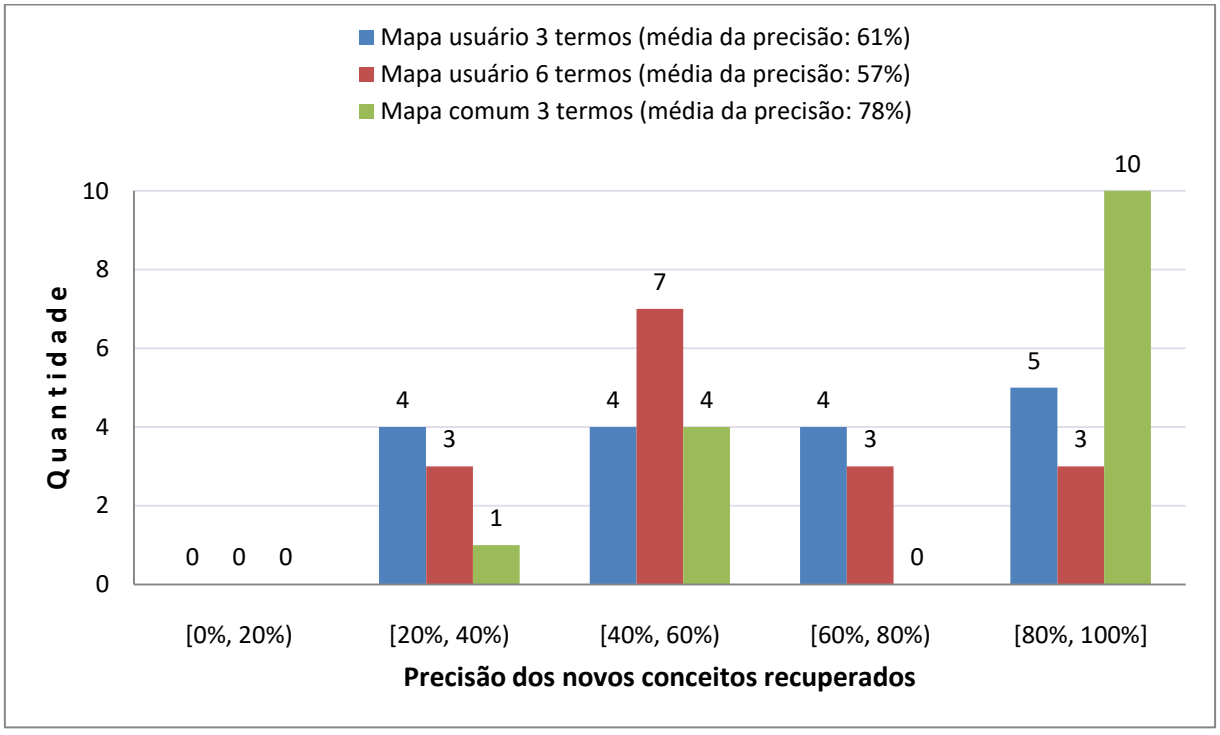
Uma análise conjunta e consolidada dos resultados mostrados no Gráfico 8, Gráfico 9 e Gráfico 10, bem como uma discussão à luz do referencial terórico são apresentadas na subseção 5.6.1 no capítulo de análise e discussão dos resultados.

4.6.3 Avaliações pelas métricas de RI

Conforme abordado na metodologia, subseção 3.5.4, o usuário preencheu o formulário de avaliação (apêndice G), questões 4(e,f,g), indicando a quantidade de conceitos e proposições relevantes, e as proposições que faltaram. A partir desses dados, baseado nas métricas de RI discutidas na subseção 2.4.5 do referencial teórico, e de acordo com o método estabelecido para medição da qualidade apresentado na subseção 3.5.5 da metodologia, foram calculados os valores relativos à:

- **Precisão dos novos conceitos recuperados**, isto é, de acordo com a subseção 3.5.5.2 da metodologia, é a razão da quantidade dos novos conceitos relevantes e recuperados sobre todos os novos conceitos recuperados;
- **Precisão das proposições recuperadas**, isto é, de acordo com a subseção 3.5.5.1 da metodologia, é a razão da quantidade das proposições relevantes e recuperadas sobre todas as proposições recuperadas;
- **Revocação das proposições recuperadas**, isto é, de acordo com a subseção 3.5.5.1 da metodologia, é a razão da quantidade de proposições relevantes e recuperadas sobre todas as proposições relevantes;

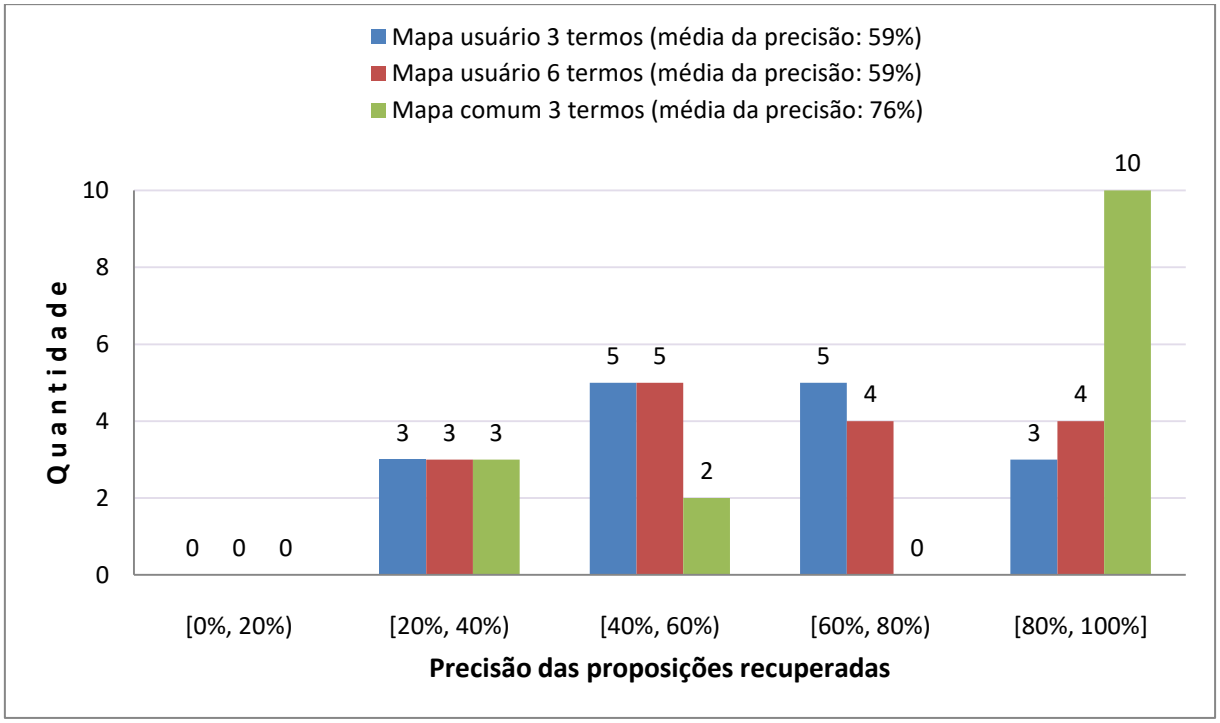
Gráfico 11 – Precisão dos novos conceitos recuperados



Fonte: Elaboração própria

O Gráfico 11 e Gráfico 12 representam a distribuição de frequência da precisão dos novos conceitos recuperados e das proposições recuperadas indicando, de forma separada, os três mapas conceituais resultantes avaliados por cada usuário. Esses gráficos mostram o resultado da precisão melhor no mapa comum de 3 termos. Os outros mapas, de 3 e 6 termos do usuário, apresentam um resultado inferior, porém, muito semelhante entre eles. Entre as duas precisões medidas nesses dois gráficos, conceitos novos e proposições recuperadas, há um resultado muito próximo, sem destaque significativo de diferença para nenhuma das medidas.

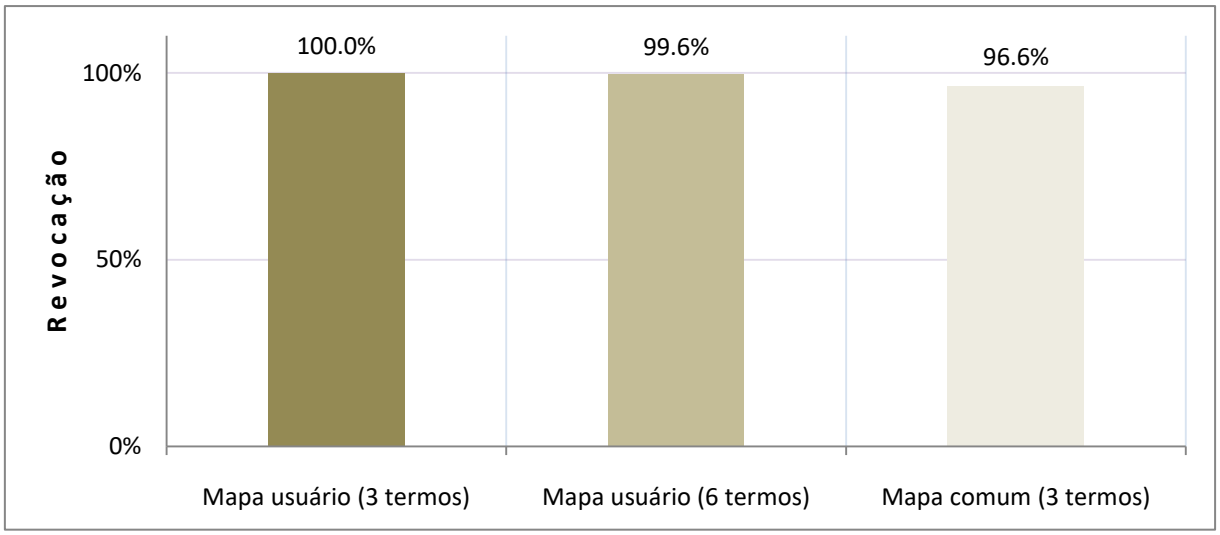
Gráfico 12 – Precisão das proposições recuperadas



Fonte: Elaboração própria

Os usuários indicaram 139 proposições faltantes ao todo nas avaliações, porém apenas 4 dessas existiam na base de conhecimento. Baseado nesse último valor e na quantidade de proposições relevantes, a média de todos os valores de revocação das proposições recuperadas foi de 99%. O Gráfico 13 representa a média dos valores de revocação das proposições recuperadas para cada um dos mapas avaliados.

Gráfico 13 – Revocação das proposições recuperadas



Fonte: Elaboração própria

Uma análise conjunta e consolidada dos resultados mostrados no Gráfico 11, Gráfico 12 e Gráfico 13, e uma discussão à luz do referencial terórico são apresentadas na subseção 5.6.2 no capítulo de análise e discussão dos resultados. Além disso, uma análise entre os resultados obtidos entre as métricas precisão e revocação são apresentados na subseção 5.6.3 do próximo capítulo.

5 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Esse capítulo faz uma análise dos resultados obtidos e os discute à luz do referencial teórico apresentado no capítulo 2. Organizado em subseções que agrupam a analise e a discussão das observações relevantes de acordo com os contextos: Ciência da Informação, recuperação de informação e conhecimento, redes complexas, mapas conceituais, experimentos sobre o modelo, avaliações realizadas pelos usuários, e base de conhecimento. Além disso, é realizada uma análise comparativa entre o modelo proposto na tese com os trabalhos correlatos apresentados na subseção 1.4.

5.1 Contexto da Ciência da Infomação

Essa seção apresenta e justifica a adoção de definições de conceito, informação e conhecimento, apresentadas no referencial terórico, que atendem quesitos e características do trabalho. Além disso, analisa e discute a equação de Brookes e os paradigmas da CI no escopo da tese.

5.1.1 Definição de conceito, informação e conhecimento adotadas

As definições de conceito propostas por Dahlberg (1978), na subseção 2.2.1 sobre Teoria do Conceito, e Novak e Gowin (1984), na subseção 2.2.3 sobre mapas conceituais, são compatíveis entre si. De um modo geral, elas descrevem conceito como sendo um enunciado sobre eventos e objetos. Observando os nós *subject* e *object* das triplas RDFs encontrados na DBpedia, constata-se que eles são descrições de eventos e objetos, tal como a definição de conceito abordada pelos autores citados. Esses mesmos nós estabelecem as redes de informação intermediárias do processo algoritmico, que depois são mapeados em conceitos no mapa conceitual resultante. Além disso, os termos de consulta fornecidos pelo usuário, que também são representados no mapa conceitual resultante, são também enquadrados nessa definição.

Sobre a definição de informação, o referencial teórico aborda uma grande diversidade delas. No presente trabalho, adota-se a definição de Brookes (1980c), subseção 2.1.3, que admite a informação como sendo um elemento que provoca transformações nas estruturas cognitivas do sujeito. Também compatível com a aprendizagem significativa de Ausubel, subseção 2.2.5, que fundamenta os mapas conceituais, pois o sujeito tem suas estruturas

cognitivas modificadas na medida em que seu conhecimento prévio estabelece relações com a nova informação, podendo transformá-la em conhecimento. De fato, no presente trabalho, o usuário recebe um conjunto de novas informações (resultado da RI) no formato de relacionamentos entre os termos de consulta do usuário, representados em um mapa conceitual. Esse mapa conceitual resultante relaciona-se com o seu conhecimento prévio, isto é, o que ele já conhece no entorno dos termos de consulta, dando-lhe condições apropriadas para a transformação de sua estrutura cognitiva.

Quanto ao conhecimento, é adotada a definição usada por Brookes (1980c) em sua equação, subseção 2.1.4. Essa definição assume o conhecimento como sendo uma estrutura de conceitos ligados por suas relações. De fato, essa definição vai ao encontro da recuperação de conhecimento adotada do modelo proposto, uma vez que os documentos recuperados são estruturados no formato de um mapa conceitual que, por sua vez, é composto por proposições que são conceitos interligados por frases de ligação, conforme é abordado na seção 2.2 do referencial teórico. Dessa forma, encarando o mapa conceitual como novo conhecimento, a outra versão da equação sugerida por Brookes, **K[S] + ΔK = K[S+ ΔS]**, absorve com mais precisão a ideia do relacionamento do novo conhecimento, ΔK, com a estrutura de conhecimento prévio do usuário, **K[S]**. Apesar de que, como o próprio Brookes argumentou, a equação continua absorvendo bem a ideia da modificação cognitiva independente da utilização de Δ**K** ou Δ**I** na equação.

5.1.2 Equação de Brookes

O Quadro 11 apresenta duas interpretações da equação de Brookes, **K[S] + ΔI = K[S+ΔS]**, para o modelo de recuperação de informação e conhecimento proposto nesse tese. Na abordagem da equação enquanto interface do usuário com o sistema (2ª coluna do Quadro 11), a estrutura cognitiva do usuário **K[S]** é modificada quando se relaciona com o mapa conceitual resultante Δ**I** provocando uma alteração em no estado do usuário Δ**S** e, consequentemente, na sua estrutura cognitiva que passa a ser representada por **K[S+ΔS]**.

Na abordagem da equação enquanto núcleo do sistema (3ª coluna do Quadro 11), a rede de informação **K[S]** formada pelas triplas RDFs recuperadas na base de conhecimento e oriundas dos termos de consulta do usuário do usuário **S**, é mesclada com a rede de informação Δ**I** formada pelos nós e ligações capazes de unificar todos os termos de consulta do usuário**.** Em seguida, essa nova rede **K[S] + ΔI** passa por um processo de transformação que provoca sua redução por intermédio de algoritmos de ranqueamento e seleção em redes

complexas até a formação do mapa conceitual resultante **K[S+ΔS]** que pode ser interpretado como uma rede de informação formada pelos termos de consulta do usuário **S** e os novos termos **ΔS** enquanto nós, e as ligações entre eles.

Quadro 11 – Interpretação da equação de Brookes no sistema de RI

Elementos equação	Abordagem enquanto interface do usuário com o sistema	Abordagem enquanto núcleo do sistema
K	Estrutura cognitiva do usuário.	Base de dados ligados.
S	Estado cognitivo prévio do usuário.	Termos de consulta do usuário.
K[S]	Estrutura cognitiva prévia do usuário.	Rede de informação formada pelas triplas RDFs recuperadas na base de conhecimento e derivadas dos termos de consulta do usuário **S**.
ΔI	Mapa conceitual resultante recebido pelo usuário.	Rede de informação com os nós e ligações capazes de unificar todos os termos de consulta do usuário da rede **K[S]**.
K[S] + ΔI	Relacionamento entre estrutura cognitiva prévia do usuário **K[S]** e o mapa conceitual resultante **ΔI**, que impulsiona a geração de uma estrutura cognitiva modificada.	Mesclagem da rede **K[S]** com a rede **ΔI**. Retorna a rede de informação expandida, com único componente conectado.
=	Transformação cognitiva, significando a geração de um novo estado mental a partir do estado prévio.	Transformação da rede, significando o impulso da rede expandida **K[S]+ΔI** para a geração de uma rede de informação reduzida (mapa conceitual) **K[S+ΔS]**
ΔS	Modificação do estado cognitivo do usuário durante a interação com o sistema.	Novos conceitos selecionados para a rede de informação reduzida (mapa conceitual).
S+ΔS	Novo estado cognitivo do	Conjunto de todos os conceitos do

	usuário, após a interação com o sistema.	mapa resultante (termos de consulta do usuário **S** e novos conceitos selecionados Δ**S**).
K[S+ ΔS]	Nova estrutura cognitiva do usuário, após o recebimento do mapa conceitual resultante, Δ**I**.	Rede de informação reduzida, mapeada no mapa conceitual resultante.

Fonte: Elaboração própria

Portanto, a partir do que foi exposto o modelo proposto tem um funcionamento compatível com a equação fundamental da CI de Brookes (1980c) - subseção 2.1.4 do referencial teórico. Tanto na visão do usuário quanto na visão interna do modelo de recuperação de informação e conhecimento proposto.

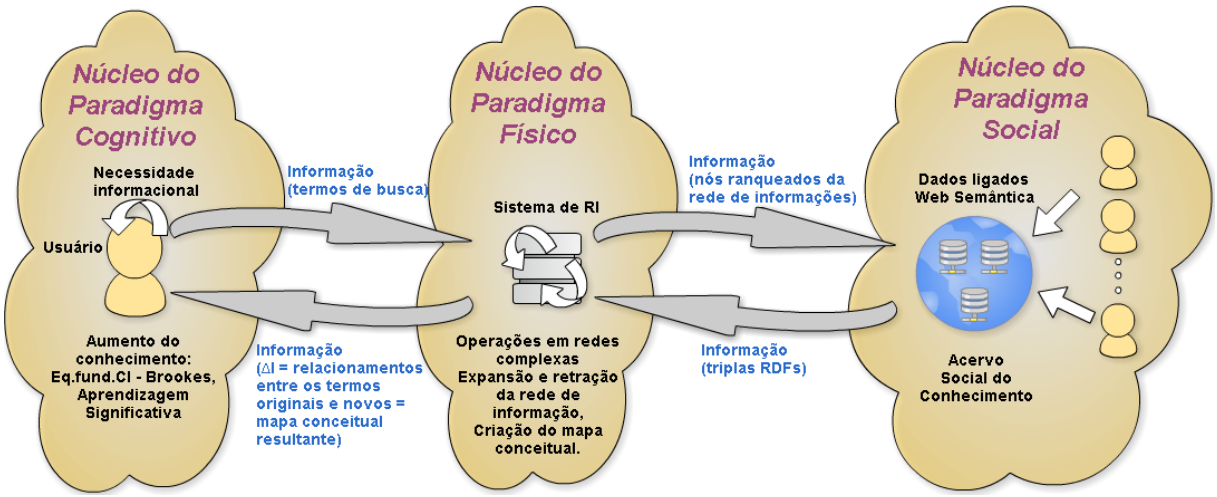
5.1.3 Paradigmas da CI

Dentre os paradigmas da CI, físico, cognitivo e social (CAPURRO, 2003), discutidos nas seções 2.1.2 e 2.4.1 do referencial teórico, o modelo de recuperação de informação e conhecimento proposto tem algumas características de cada um deles. Essa abordagem simultânea dos três paradigmas está em consonância com a visão defendida por Araujo (2014) quando ele diz que os "[...] problemas informacionais continuam tendo uma dimensão física, tendo também aspectos cognitivos e se inserindo em dimensões contextuais e pragmáticas". Apesar de adotar os três paradigmas, por outro lado, o modelo descarta características desfavoráveis, tais como, o distanciamento do usuário no paradigma físico, a isenção do aspecto social no paradigma cognitivo e a despreocupação com a tecnologia inerente ao paradigma social. A Figura 78 representa o modelo geral contextualizado nesses três paradigmas da CI.

O **núcleo do paradigma cognitivo**, na Figura 78, é composto pelo usuário e seus processos cognitivos, a representação de sua necessidade informacional por intermédio dos termos de consulta do usuário, e o recebimento da informação recuperada representada pelo mapa conceitual resultante. O paradigma cognitivo representa a dimensão informacional no processo e pode-se dizer que ele segue uma abordagem subjetiva da informação (ARAÚJO, 2014) sendo admitido pelo modelo proposto no início e no término de todo o processo da recuperação de infomação e conhcimento. Na parte inicial do processo, quando o indivíduo constata um estado anômalo de conhecimento, ele formula o conjunto de termos para

representar sua necessidade informacional e o fornece ao sistema de RI. A parte final do processo é baseada na equação fundamental da CI de Brookes (1980c), **K[S] + ΔI = K[S+ΔS]**, conforme discutido na subseção 5.1.2, Quadro 11 - abordagem enquanto interface do usuário com o sistema. Os mapas conceituais (NOVAK, 1977), fundamentados na teoria da aprendizagem significativa (AUSUBEL, 1968), subseção 2.2.5 do referencial teórico, representam os relacionamentos existentes entre os termos de consulta do usuário resultantes do processamento do sistema de RI. Assim, o usuário recebe o mapa conceitual resultante **ΔI** que, relacionando-se com a sua estrutura cognitiva prévia **K[S]**, provoca mudanças **ΔS**, em seu estado cognitivo **S**, provocando também mudanças em sua estrutura cognitiva, passando a ser representada na equação pelo termo **K[S+ΔS]**.

Figura 78 – Modelo geral de RI contextualizado nos paradigmas da Ciência da Informação



Fonte: Elaboração própria

O núcleo do paradigma físico, na Figura 78, é composto pelo sistema de RI e se encarrega de grande parte do aspecto tecnológico do modelo, englobando desde a leitura de dados na base de conhecimento, passando por todas as operações em redes complexas que fazem parte da expansão e da retração da rede de informação, e também o processo de mapeamento para criar o mapa conceitual resultante. O paradigma físico, apesar de ter sido o primeiro modelo adotado pela CI, ainda mantém sua atualidade, sobretudo na construção dos motores de busca na internet (ARAÚJO, 2014).

O núcleo do paradigma social, na Figura 78, é composto pela base de conhecimento de dados ligados na Web Semântica e todo movimento social pra a sua criação e manutenção. O paradigma social se apresenta como uma abordagem sócio-cognitiva, levando em

consideração o conhecimento compartilhado por uma comunidade ou grupo (ALMEIDA *et al.*, 2007) e com foco na informação construída (NASCIMENTO, 2006). O acervo social do conhecimento, representado pela base de dados ligados, é fundamentada numa construção coletiva obedecendo a domínios bem estabelecidos por intermédio de suas ontologias. O acervo é dinâmico e cresce segundo os preceitos de movimentos tais como o *crowdsourcing* (subseção 2.3.1 do referencial teórico).

5.2 Contexto da recuperação de informação e conhecimento

Essa seção analisa e discute, baseado no referencial terórico, o porquê da adoção de modelo híbrido para o trabalho. Também apresenta como ocorre a recuperação de conhecimento, a visualização de informação e conhecimento, e a visualização de domínio do conhecimento no trabalho.

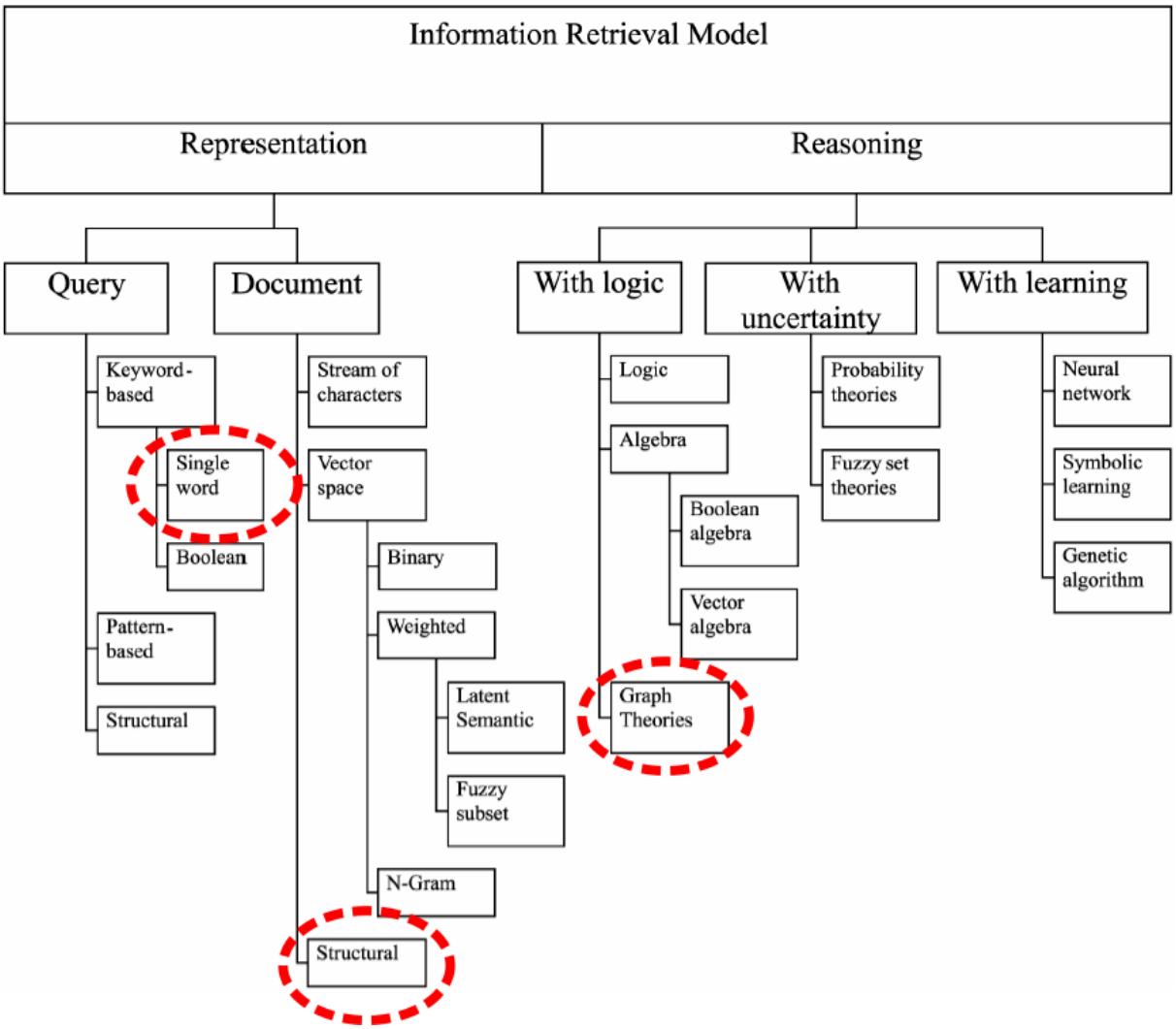
5.2.1 Modelo híbrido

Com o objetivo de produzir respostas úteis aos usuários, os sistemas modernos de RI possuem características de vários modelos ao invés de se enquadrarem em apenas um tipo (BAEZA-YATES; RIBEIRO-NETO, 2011). A partir das taxonomias apresentadas na subseção 2.4.7, observou-se que a presente proposta de modelo de RI se enquadra em alguns itens das taxonomias de Canfora e Cerulo (2004) e Baeza-Yates e Ribeiro-Neto (2011). Esses itens das taxonomias são indicados de forma destacada na Figura 79, Figura 80 e Figura 81.

A justificativa do enquadramento do modelo proposto em cada um dos itens destacados nessas figuras é explicada nos tópicos a seguir:

- ***Representation - query - single word*** (Figura 79): a representação da consulta acontece por intermédio de palavras simples, uma vez que a entrada do usuário é fornecida como uma lista de termos;
- ***Representation – document – structured*** (Figura 79): a representação do documento é de forma estruturada, pois a base de conhecimento é formada por RDFs;
- ***Reasoning – with logic – graphic theories*** (Figura 79): o método de condução da recuperação é por intermédio de lógica e com uso da teoria de grafos, isto é, as métricas de redes usadas no ranqueamento e seleção de nós da rede de informação têm como base a teoria dos grafos.

Figura 79 – Enquadramento do modelo proposto na taxonomia vertical



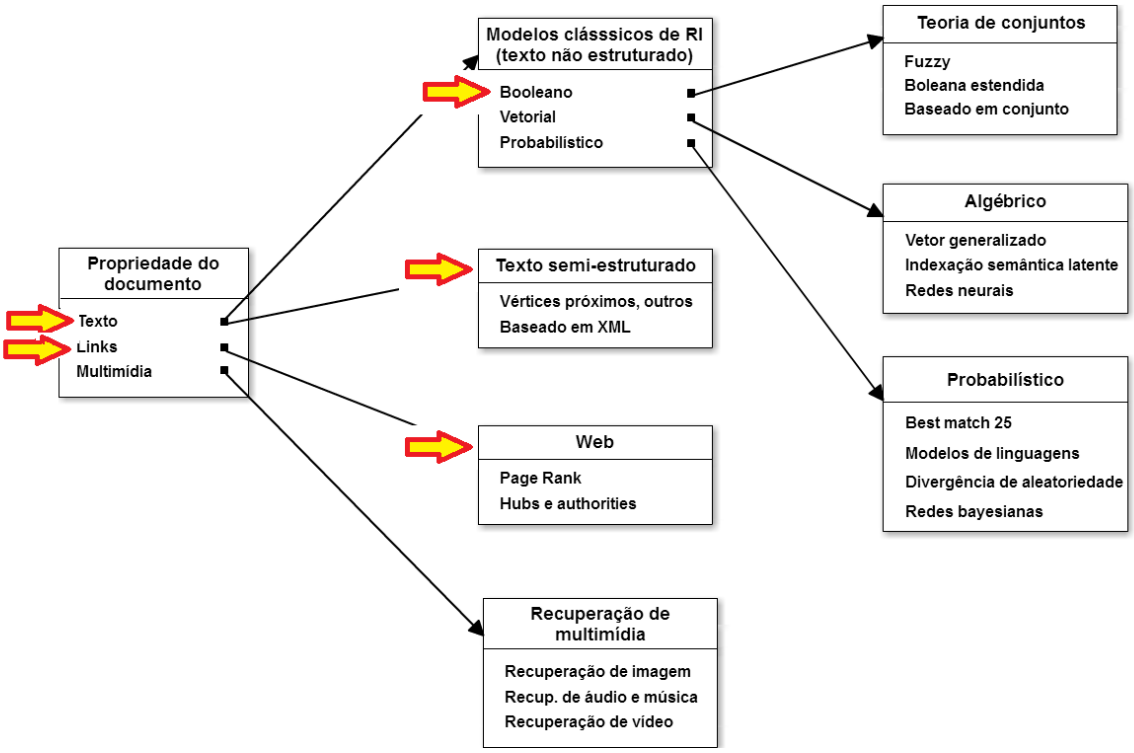
Fonte: adaptado de Canfora e Cerulo (2004, p. 177)

- **Texto – booleano** (Figura 80): a busca por termos textuais, mesmo que numa base estruturada em RDFs, porém formada por elementos textuais, caracteriza-se por uma consulta booleana de forma restrita, onde apenas o operador de conjunção, OR, é usado, pois basta que exista algum dos termos fornecidos pelo usuário para que a tripla RDF seja recuperada;
- **Texto – semi-estruturado** (Figura 80): a estrutura em triplas RDFs, onde cada um dos seus elementos é textual, pode ser caracterizada como um texto estruturado;
- ***Link* – Web – *Page Rank*** (Figura 80): a base de dados ligados por si só já caracteriza o modelo proposto com a propriedade “link” onde os documentos, triplas

RDFs, estão ligados entre si. Um dos algoritmos usados no processo de ranqueamento é o *eigenvector*, que serviu de base algorítmica para a criação do método *Page Rank*.

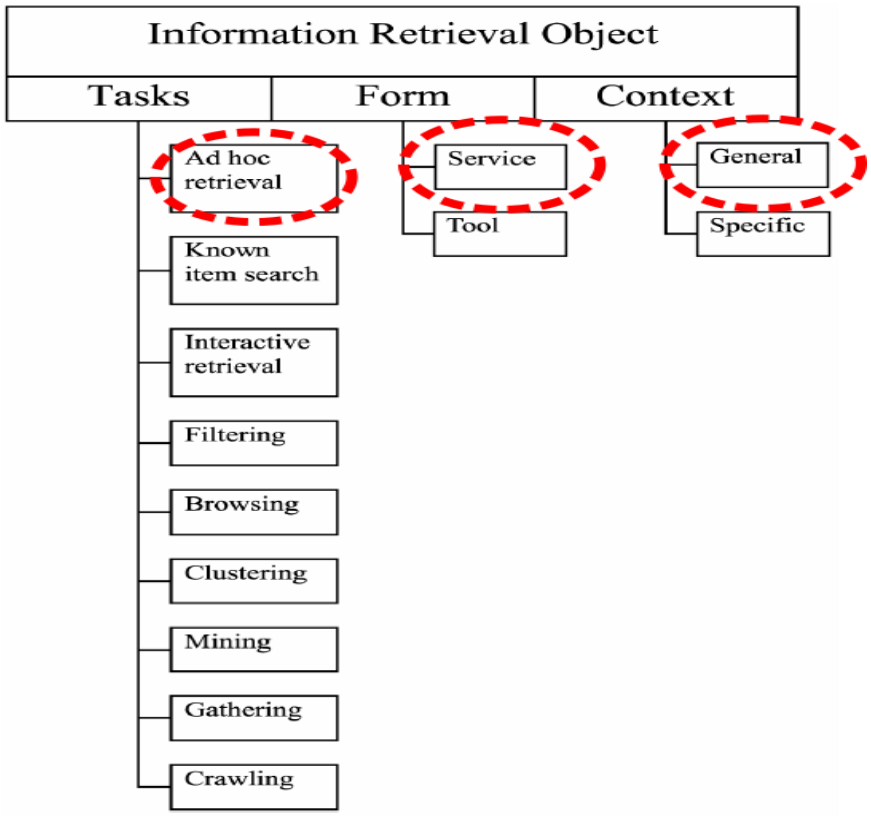
- ***Link – Hub* e *Authorities*** (Figura 80): além do algoritmo *eingenvector* (citado no item anterior) são usados também métricas tais como *betweenness*, *closeness*, *eccentricity*, quantidade de componentes conectados que, de certa forma, têm relações com *hub* e *authorities*.
- ***Tasks - ad hoc retrieval*** (Figura 81): as consultas são independentes sem depender da continuidade de interatividade com o usuário, ou seja, o usuário fornece o conjunto de termos e obtém o mapa conceitual resultante sem interações no meio do processo;
- ***Form - service*** (Figura 81): a sua forma de atuação é classificada como 'serviço', pelo fato de oferecer o serviço de busca que é entregue via web;
- ***General*** (Figura 81): trabalha num domínio de conhecimento geral, pois utiliza a base de dados ligados de forma completa.

Figura 80 – Enquadramento do modelo proposto na taxonomia completa



Fonte: adaptado de Baeza-Yates e Ribeiro-Neto (2011, p. 60)

Figura 81 – Enquadramento do modelo proposto na taxonomia horizontal



Fonte: adaptado de Canfora e Cerulo (2004, p. 184)

As duas taxonomias apresentadas aqui no enquadramento do modelo e também as outras estudadas na seção 2.4.7 do referencial teórico, não são suficientes para classificar de maneira adequada o modelo proposto de RI. Existe um componente forte no modelo, relativo à topologia dos documentos, ou seja, a forma como eles se ligam. Esse componente foi fundamental para a recuperação dos conceitos da rede informacional formada durante o processo de RI. Além disso, existe uma retroalimentação de termos no processo que é também fundamental para a formação dessa rede informacional e, consequentemente, na melhoria da recuperação dos conceitos para a formação do mapa conceitual final. Considerando esses fatores e também o seu enquadramento em várias categorias como mostrado na Figura 79, Figura 80 e Figura 81, o modelo proposto foi categorizado como híbrido.

5.2.2 Recuperação de conhecimento

O modelo proposto também pode ser enquadrado enquanto recuperação de conhecimento, baseado nos critérios comparativos estabelecidos por Van Rijsberngen (1979)

e estendidos por Yao *et al.* (2007), conforme discutido na subseção 2.4.3 do referencial teórico. O Quadro 12 mostra em destaque os elementos usados no modelo proposto, que em maior número estão na coluna da recuperação de conhecimento.

Quadro 12 – Identificação de características de recuperação no modelo proposto

	Recuperação de dados	Recuperação de informação	Recuperação de conhecimento
Casamento da busca	booleano	parcial ou melhor	parcial ou melhor
Inferência	dedutiva	indutiva	dedutiva, indutiva, raciocínio associativo, raciocínio analógico
Modelo	determinístico	estatístico e probabilístico	semântico, de inferência
Consulta	linguagem artificial	linguagem natural	estrutura de conhecimento, linguagem natural
Organização	tabela, índice	tabela, índice	unidade de conhecimento, estrutura de conhecimento
Representação	número, regra	linguagem natural, linguagem de marcação	grafo de conceitos, lógica de predicados, regra de produção, estrutura, rede semântica, ontologia
Armazenamento	banco de dados	coleções de documentos	base de conhecimento
Resultados recuperados	conjunto de dados	seções ou documentos	um conjunto de unidade de conhecimento

Fonte: adaptado de Yao *et al.* (2007) extensão de Van Rijsberngen (1979)

Seguem as justificativas para o enquadramento do aspecto enquanto recuperação de dados, de informação ou de conhecimento:

- **Casamento de busca**: é através de comparação booleana quando um determinado termo do usuário é buscado na base de dados ligados;
- **Inferência**: é dedutiva, pois para se chegar aos resultados é preciso percorrer um caminho preciso e lógico;

- **Modelo**: é semântico, pois a recuperação usa ligações da rede de informação formada por conceitos;
- **Consulta**: usa a linguagem artificial SPARQL;
- **Organização**: usa unidade de conhecimento, que são os RDFs das bases de dados ligados;
- **Representação**: é feita por grafos de conceitos (redes de informação), redes semânticas (mapas conceituais) e ontologias nas bases de dados ligados;
- **Armazenamento**: é por intermédio de base de conhecimento, representada por bases de dados ligados;
- **Resultados recuperados**: é um conjunto de unidades de conhecimento, representado pelo mapa conceitual resultante.

5.2.3 Visualização de informação e conhecimento

Na etapa final do processo de recuperação de informação e conhecimento, o modelo apresenta ao usuário um mapa conceitual representando uma estrutura que contém relacionamentos entre os termos de consulta do usuário. Conforme discutido no referencial teórico, subseção 2.4.4.1, a visualização de estruturas é a base para a visualização de informação e conhecimento (KELLER; TERGAN, 2005). A visualização de conhecimento foca na transferência e criação de conhecimento (BURKHARD, 2005; MEYER, 2010) que, no caso do modelo proposto, é representado pelo mapa conceitual resultante. Além disso, diagramas conceituais são exemplos de visualização de conhecimento, pois eles são descrições esquemáticas de ideias abstratas com o auxílio de formas padronizadas usadas para estruturar a informação e ilustrar relacionamentos (EPPLER; BURKHARD, 2004). Dadas as características de um mapa conceitual, ele é também considerado um diagrama conceitual.

Na subseção 2.4.4.2 do referencial teórico, Zhang (2008) destaca o quanto a visualização de informação pode ser benéfica na RI. Especificamente para o modelo proposto e considerando o mapa conceitual para apresentar a informação recuperada, destacam-se os seguintes benefícios:

- Realiza a espacialização de um contexto informacional, isto é, torna um espaço de informação invisível e abstrato para um espaço visual;
- Oferece caminhos diferenciados para o desenvolvimento de novos modelos de RI, diferentes dos modelos tradicionais;

- Fornece condições bastante favoráveis para a análise da informação, como por exemplo, a análise de conexões entre informações;
- Abre amplas possibilidades para o desenvolvimento de abordagens para as apresentações visuais, principalmente pela espacialidade; e
- Enriquece e eleva o nível da RI tornando o processo intuitivo e simples, e ainda deixando os usuários capazes de construir e realizar descobertas de conhecimento.

O Quadro 3, da subseção 2.4.4.1 do referencial teórico, faz uma comparação entre visualização de informação e visualização de conhecimento (BURKHARD, 2005). Entre os aspectos citados por Burkhard, destacam-se alguns com características que se aproximam do modelo proposto quanto ao enquadramento enquanto visualização de conhecimento, basicamente pelo uso de mapas conceituais na apresentação da informação recuperada:

- Objetivo: uso de representações visuais para melhorar a transferência de conhecimentos entre as pessoas e melhorar a criação de conhecimento em grupos;
- Benefício: multiplica os processos de transferência de conhecimento e de comunicação entre os indivíduos usando uma ou mais representações visuais;
- Destinatários: auxilia um indivíduo ou um grupo na transferência ou criação de novos conhecimentos em ambientes colaborativos;

Ainda na subseção 2.4.4.1 do referencial teórico, Eppler e Burkhard (2004) destacam vantagens da visualização de conhecimento, sobre a visualização de informação. Entre as vantagens e pela proximidade com o modelo proposto, destacam-se os seguintes benefícios:

- Benefício social: coordenação na comunicação entre profissionais do conhecimento;
- Benefício cognitivo: sensibilização e foco para a criação e transferência do conhecimento; maior compreensão e apreciação de conceitos e ideias e suas relações; e revelação de conexões escondidas.

Portanto, o modelo proposto, na etapa da apresentação da informação e conhecimento recuperados, por intermédio do mapa conceitual resultante, pode ser enquadrado enquanto visualização de informação, contudo, o tipo de visualização predominante, é a visualização de conhecimento.

5.2.4 Visualização de domínio de conhecimento

A visualização de domínio de conhecimento é tratada na subseção 2.4.4.3 do referencial teórico. Basicamente, é uma visualização com o objetivo de mostrar os detalhes estruturais e as características mais importantes de um determinado domínio de conhecimento (HOOK; BÖRNER, 2005), sendo que esse conhecimento, algumas vezes, precisa ter sua fronteira descoberta ou revelada (CHEN, 2013). Além disso, segundo Hook e Börner (2005), a visualização de domínio de conhecimento pode ser considerada um mapa de domínio.

Dessa forma, e considerando a discussão sobre a visualização de conhecimento oferecida pelo modelo proposto, na subseção 5.2.3, os mapas conceituais do modelo, que apresentam o conhecimento recuperado, podem também auxiliar a delimitação de um domínio do conhecimento, isto é, se os termos de consulta do usuário pertencerem a esse mesmo domínio. Se o número de conceitos intermediários for minimizado na configuração do sistema de RI, então os relacionamentos, os novos conceitos intermediários e os termos de consulta do usuário ficarão dentro do mesmo domínio, facilitando assim, a descoberta, a delimitação e a visualização desse domínio de conhecimento.

5.3 Contexto de redes complexas

Baseado no referencial teórico, essa seção analisa e discute alguns aspectos, propriedades e fenômenos de redes complexas que ocorreram no trabalho. Algumas observações ocorreram em duas etapas distintas: (1) **rede intermediária**: é a rede que vai se formando desde o início do algoritmo até atingir uma fase onde todos os termos do usuário estão conectados num único componente gigante, antes de iniciar o processo de redução, e atingindo, na maior parte das vezes, seu ápice de tamanho, e (2) **rede final**: é a rede de informação final após o processo de redução, mapeada no mapa conceitual resultante.

5.3.1 Topologia de rede não direcionada

No experimento inicial, descrito na seção 4.4, observou-se que a topologia direcionada das redes intermediárias, criadas desde o mapeamento do conjunto de RDFs até a rede final mapeada no mapa conceitual resultante, não foi adequada para algumas medições. Por exemplo, o relacionamento ‘Paul Otlet influenciou Tim Berners-Lee’ não deveria ser

diferente, ou ter mais peso, do relacionamento inverso ‘Tim Berners-Lee foi influenciado por Paul Otlet’.

O foco do modelo de RI proposto não é o significado semântico de uma ou outra relação, mas o quanto um nó ou uma relação são importantes perante a rede como um todo e numa visão topológica total, ou seja, como eles estão dispostos ao longo da rede ou o quanto eles influenciam um possível fluxo informacional.

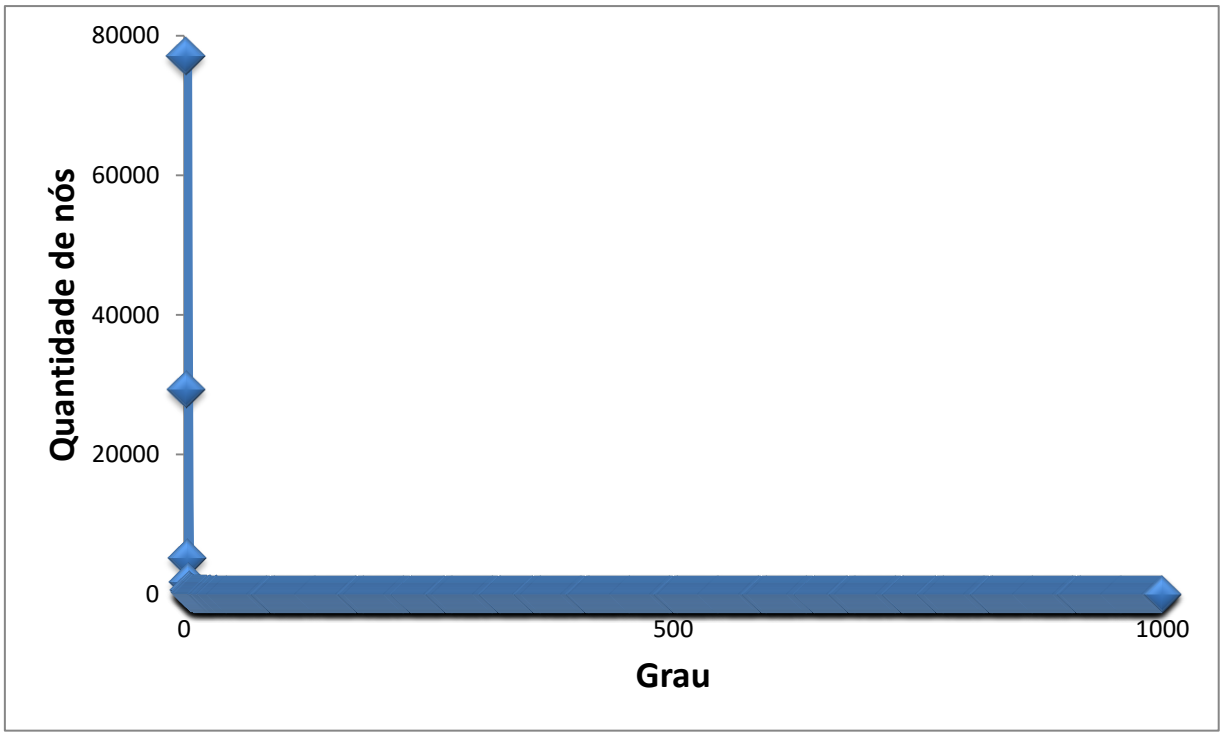
5.3.2 Consulta completa para formação da rede

Newman (2010) alerta que numa análise de rede não se pode usar amostragem, mas, se for o caso, deve-se diminuir a questão geradora dos dados para diminuir a quantidade de informações coletadas para a rede. Assim, o modelo proposto usa todas as triplas RDFs coletadas pelas consultas formando uma rede informacional total. Em algumas situações a rede de informações intermediária fica com tamanho grande. Porém, a rede algumas vezes é reduzida, mas, isto ocorre em função de algumas operações de ranqueamento que descartam o número excessivo de nós, desprezando aqueles com menor importância, tal como mostrado no experimento da subseção 4.4, e no algoritmo do modelo aprimorado na subseção 4.5.4. Além disso, um fator também considerado para a eliminação do excesso de nós de uma rede é o custo computacional que pode se tonar grande em redes grandes.

5.3.3 Rede livre de escala ou scale-free

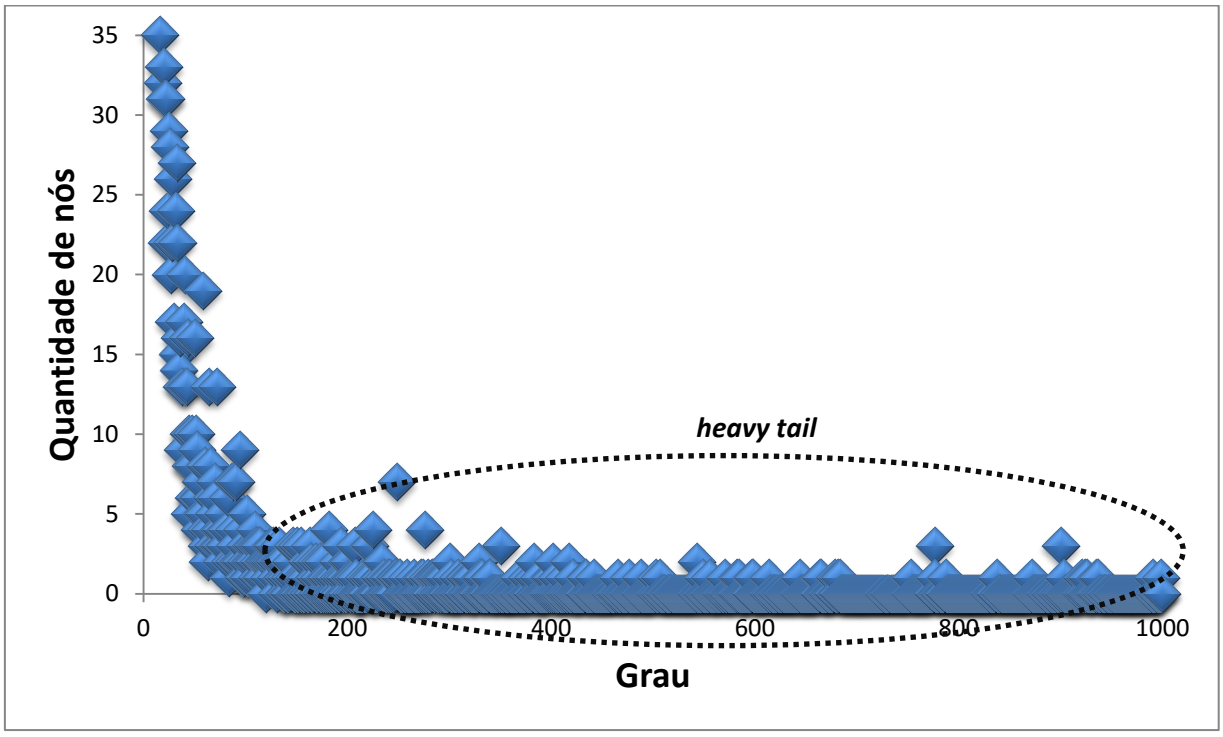
Observou-se nas 32 redes intermediárias usadas na validação o fenômeno livre de escala ou *scale-free*, abordado na subseção 2.5.2.4 do referencial teórico, onde a distribuição dos nós acontece pela lei de potência. A observação foi visual com auxílio do software Gephi comparando cada um dos gráficos gerados pelas redes intermediárias com o aspecto do Gráfico 5 (subseção 2.5.2.4) que representa a distribuição de potência. Para facilitar a visualização aqui no presente trabalho, foram condensadas as 32 redes analisadas num único gráfico, apenas juntando-se todos os nós e seus respectivos graus. O resultado pode ser conferido no Gráfico 14.

Gráfico 14 – Distribuição de frequência do grau dos nós das 32 redes intermediárias



Fonte: Elaboração própria

Gráfico 15 – Distribuição de frequência do grau dos nós das 32 redes intermediárias, com ampliação da visualização e destaque para o *heavy tail*



Fonte: Elaboração própria

Devido ao número muito alto da quantidade de nós com baixo grau, da ordem de 80 mil, foi feita uma redução de escala no eixo Y-vertical do gráfico que representa essa quantidade para ampliar a visualização próxima à interseção dos eixos, Gráfico 15. Nesse gráfico, a presença do *heavy tail* (discutido na subseção 2.5.2.4) é bem aparente, representando os poucos nós com grande quantidade de conexões.

Estudos apresentados na subseção 2.5.2.4 afirmaram que a web é livre de escala. As redes intermediárias analisadas são um subconjunto dos dados ligados da DBpedia, e estes são um subconjunto de informações da Wikipedia que, por sua vez, é um subconjunto da web. Porém, não é possível afirmar que as redes intermediárias são livres de escala devido a existência dessa associação de composição.

Quanto ao modelo proposto no presente trabalho, observa-se que a presença de *hubs* facilita o algoritmo no processo de redução da rede com mais determinismo, uma vez que a escolha dos elementos de maior excentricidade, para eliminação de nós na rede, mantém os nós importantes, como os *hubs*, para a formação dos mapas conceituais resultantes.

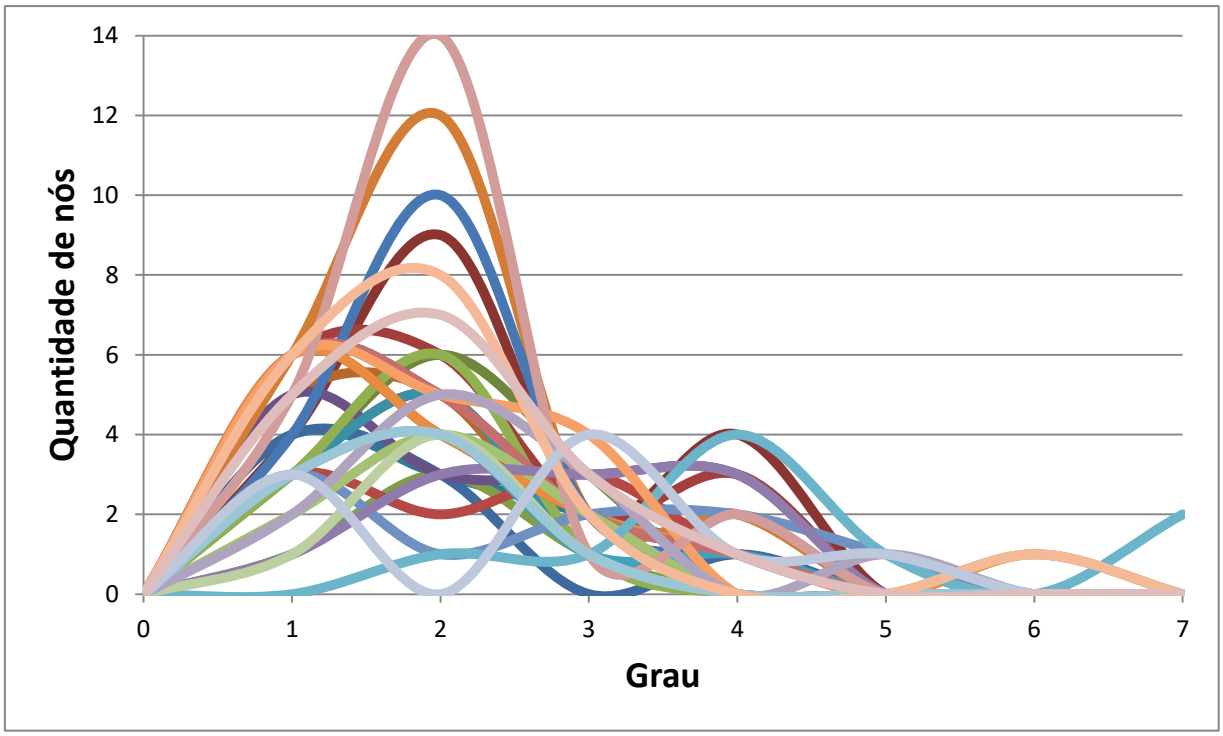
5.3.4 Rede aleatória

Nas redes finais não foi observado o fenômeno livre de escala, como aconteceu nas redes intermediárias da subseção 5.3.3. Contudo, observou-se que as distribuições tiveram um leve aspecto de Poisson (curva do sino representada pelo Gráfico 4 na subseção 2.5.2.3 do referencial teórico), como mostra o Gráfico 16, que representa as 32 redes finais. Apesar das redes aleatórias possuírem a forte caracterísitica de serem geradas por processos computacionais aleatórios, observou-se que a distribuição de frequência de graus de nós das redes finais, da validação realizada no presente trabalho, tiveram curvas muito semelhantes à distribuição de Poisson, tipicamente encontrada em redes aleatórias.

5.3.5 Rede mundo pequeno ou small-world

Quando acontece da média de todos os caminhos mínimos de uma rede (*average path length*) possuir um valor muito pequeno em comparação a sua quantidade de nós, é um forte indício de que ela possui o fenômeno mundo pequeno ou *small world*, abordado na subseção 2.5.2.5 do referencial teórico.

Gráfico 16 – Distribuição de frequência do grau dos nós das 32 redes finais



Fonte: Elaboração própria

A Tabela 1 apresenta a quantidade de nós e a média dos caminhos mínimos das 32 redes intermediárias analisadas, sendo possível observar uma diferença proporcional grande entre os dois valores medidos em cada rede. A média da quantidade de nós é 3.644 e a média de todas as médias dos caminhos mínimos é 3,79.

Tabela 1 – Relação entre quantidade de nós e média do caminho mínimo das redes intermediárias

Qtde total de nós	Média caminho mínimo
24.599	3,44
11.316	2,34
10.258	3,13
8.570	3,64
6.647	3,64
6.237	3,37
6.023	4,45
5.816	4,53
5.587	3,68
5.474	5,26
5.234	3,42

Qtde Total de nós	Média caminho mínimo
5.068	4,27
4.674	3,34
4.550	4,46
3.987	3,88
3.677	2,60
3.627	4,30
3.205	4,56
2.991	4,07
2.877	3,71
2.826	4,15
2.815	4,00

Qtde total de nós	Média caminho mínimo
2.774	4,02
2.641	3,41
2.210	3,37
2.164	4,25
2.090	2,96
2.085	3,99
2.036	4,10
1.137	3,88
799	3,79
301	3,29

Fonte: Elaboração própria

Também foram observados, nessas redes intermediárias, valores de diâmetro baixo e alto coeficiente de clusterização, que são características típicas de redes com o fenômeno mundo pequeno, tal como abordado no referencial teórico subseção 2.5.2.5.

Quanto às redes finais, que representam os mapas conceituais resultantes, a média da quantidade de nós é 12 e a média de todas as médias dos seus caminhos mínimos é 2,98. Nesse caso, a verificação de tal fenômeno não é relevante, pois as redes são pequenas.

O fenômeno mundo pequeno identificado nas 32 redes intermediárias construídas a partir de três ou seis termos fornecidos pelos usuários representa o quanto os conceitos estão próximos um dos outros mesmo sendo de áreas distintas, como aconteceu em algumas situações na coleta de dados. Portanto há indícios da existência de possibilidade de ligação para quaisquer conjuntos de termos fornecidos e com poucos conceitos intermediários entre eles, isto é, considerando-se a base de conhecimento usada na validação: DBpedia.

5.3.6 Rede descentralizada

Foi observada a presença de *hubs* nas redes intermediárias. Essa observação é compatível com a ocorrência do fenômeno livre de escala (subseção 5.3.3), implicando em mais uma indicação da presença de poucos nós com grau elevado. A presença desses *hubs* é também um forte indicativo de que as redes intermediárias não são distribuídas, mas descentralizadas (tal como abordado na subseção 2.5.2.2 do referencial teórico). Um exemplo de rede intermediária com esse aspecto pode ser conferido na seção 4.5.5, referente a uma execução do protótipo, da Figura 58 a Figura 61.

5.3.7 Rede distribuída

A partir do referencial teórico, subseção 2.5.2.2, observou-se que as redes finais dos testes realizados no modelo e da validação, não são centralizadas e nem descentralizadas, mas, distribuídas. É possível observar esse aspecto topológico nos mapas conceituais resultantes da validação com os usuários que estão disponíveis no APÊNDICE H. O aspecto de topologia distribuída nas redes finais gera mapas conceituais mais interessantes pelo fato dos conceitos se interligarem de forma mais homogênea e distribuída com os outros conteitos. Mapas com conceitos mais centralizadores diminuem a integração mais abrangente de conceitos advindos de áreas diferentes e possuem menos *Crosslinks* (subseção 2.2.3 do referencial teórico).

5.3.8 A força dos laços fracos ou the strength of weak ties

A força dos laços fracos, ou *the strength of weak ties*, tal como abordado na subseção 2.5.2.6 do referencial teórico, foi um fenômeno observado no funcionamento do modelo proposto nos testes piloto e na validação com os usuários. Os nós que faziam a intermediação entre as várias subredes coesas, foram determinantes para auxiliar a agregação dos termos fornecidos pelo usuário num único componente gigante. Todavia, esse fenômeno só foi observado no crescimento das redes intermediárias, pois enquanto as redes crescem há demanda por nós de intermediação para capturar ligações com outras subredes diminuindo a quantidade de componentes conectados. Um exemplo pode ser conferido na rede da seção 4.5.5 referente à execução do protótipo, principalmente na Figura 56 e Na primeira rede de informação, Figura 56, dois termos do usuário, 'Sociology' e 'Social network', conseguiram conexão entre si, e o nó proveniente do termo 'Liquid modernity' encontra-se em um componente conectado isolado dos demais. Somente na terceira iteração, Figura 58, é que ele consegue conexão com os demais nós formando único componente conectado, sendo mostrado no log do Quadro 8 como '*** Iteration 2 ... Connected components: 1', e tendo a rede crescido de 657 nós para 1.692 nós.

Figura 57, onde há ocorrência de nós intermediários entre as várias subredes.

5.3.9 Aptidão ou fitness

A aptidão ou *fitness*, tal como abordado na subseção 2.5.2.8 do referencial teórico, pode ser observada no crescimento da rede de informação quando, no meio do processo, é descoberto um nó com grau muito grande. Esse nó, mesmo secundário e não sendo importante na ligação dos termos de consulta do usuário, consegue atrair a descoberta de mais nós ligados a ele devido a sua grande interface com o restante da rede de dados ligados. Por exemplo, no experimento, o nó denominado 'United States', sozinho, tem um grau maior que 10 mil, muito acima que a maioria.

A ocorrência desse fenômeno no modelo proposto, poderia se tornar um problema, pois com alguns nós de aptidão diferenciada na rede poderia fazê-la assumir um tamanho grande e inviável para a execução dos algoritmos de ranqueamento devido ao alto custo computacional. Contudo, quando isso acontece, o processo dispara a aplicação de filtros de grau dois e a execução do algoritmo 2-core em pontos estratégicos do algoritmo, como pode

ser observado no algoritmo do modelo aprimorado na subseção 4.5.4. Assim, grande parte dos nós coletados na base de conhecimento, em função da grande aptidão de poucos, é descartada para tornar as operações de rede com um custo computacional controlado.

5.4 Contexto dos mapas conceituais

Essa seção analisa e discute três aspectos de mapas conceituais que são importantes no presente trabalho: o quanto os mapas são úteis para apresentar informações; os mapas conceituais não tem obrigatoriedade de hierarquia conceitual; e sobre a quantidade total de conceitos.

5.4.1 Mapa conceitual para apresentação da informação recuperada

Os mapas conceituais são boas ferramentas para representação de conhecimento, conforme apresentado na subseção 2.2.3 sobre a sua caracterização, e na subseção 2.2.6 sobre aplicações diversas. Os mapas conceituais são também boas ferramentas para apresentação e disseminação de informações, conforme discutido na subseção 2.2.7 do referencial teórico, tendo confirmações nesse sentido sido realizadas por autores, tais como Vekiri (2002), Lima (2004), Valerio, Leake e Cañas (2012), Orrantia (2012), Conceição, Samuel e Biniecki (2014), e Cañas *et al.* (2015).

Soma-se a esses resultados, o fato do mapa conceitual ser considerado uma boa ferramenta para proporcionar a visualização de conhecimento (*knowledge visualization*), tal como discutido na subseção 2.4.4 do referencial teórico e na subseção 5.2.3 desse capítulo.

Dessa forma, conclui-se que o mapa conceitual constitui-se numa boa abordagem para apresentação da informação recuperada.

5.4.2 Mapa conceitual sem a obrigação de hierarquia

O modelo proposto assume que mapas conceituais não necessitam de hierarquia. Tal como já discutido nas considerações finais do referencial teórico, subseção 2.6, alguns pesquisadores encararem os mapas conceituais como estruturas hierárquicas (MOREIRA; MASINI, 1982; SHERRATT; SCHLABACH, 1990; CAÑAS; NOVAK; REISKA, 2015). Contudo, Dutra, Fagundes e Cañas (2004) argumentam que são as frases de ligação que estabelecem as fronteiras de um conceito no mapa. Lanzing (1997) argumenta que mapa

conceitual é uma rede de conceitos. Portanto, como as redes consistem de nós e ligações sem restrição hierárquica, com possibilidade livre para conexões, e com o intuito de abrir o leque de possibilidades para as fronteiras de um conceito, o modelo proposto adota essa topologia livre para os mapas conceituais a fim de representar com mais fidelidade a rede informacional final gerada pelo processamento que é depois mapeada no mapa conceitual resultante da RI. Dessa forma, o mapa conceitual resultante não tem foco na hierarquia e sim nas proposições e principalmente na revelação de elementos intermediários que fazem a ligação dos termos originalmente fornecidos pelo usuário.

5.4.3 Quantidade de novos conceitos no mapa conceitual

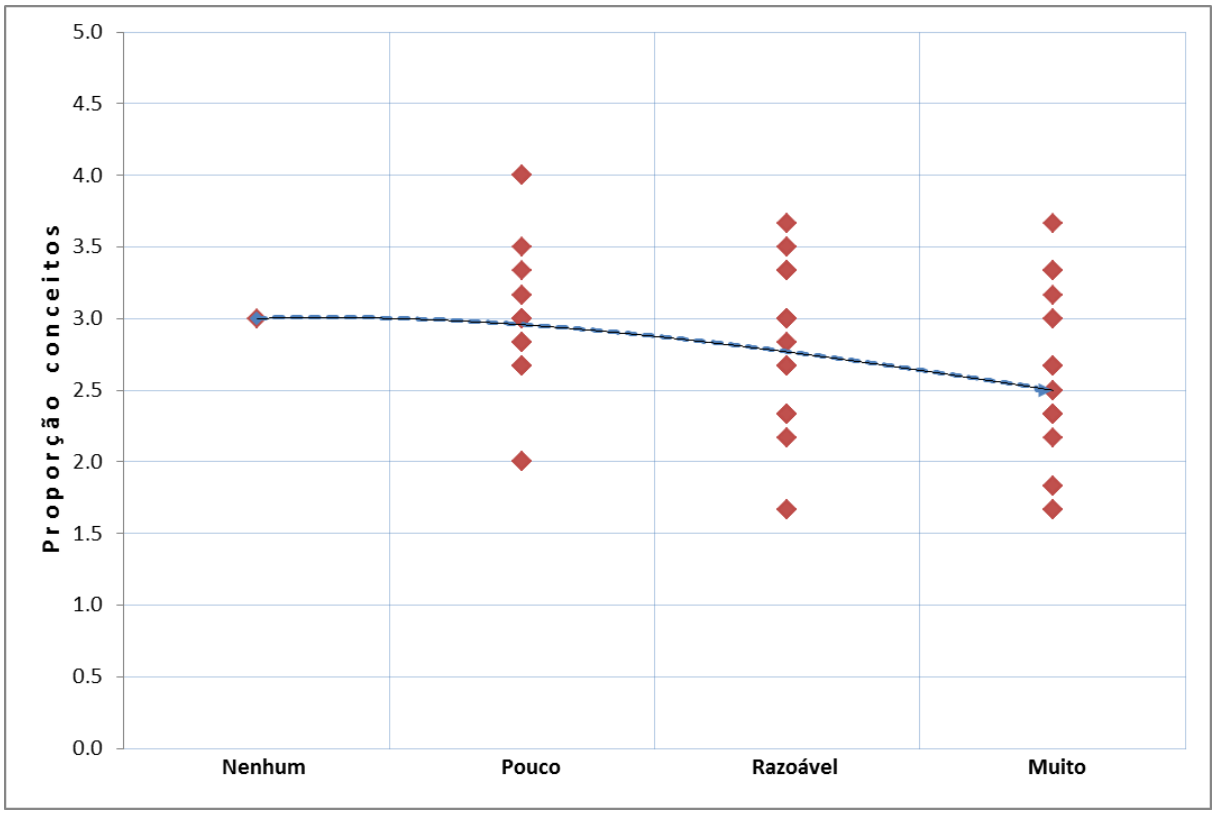
A quantidade de conceitos novos nos mapas conceituais resultantes varia em função da necessidade do algoritmo em realizar mais iterações e, assim, encontrar mais elementos intermediários para unir todos os termos do usuário num único componente conectado. Dessa forma, alguns mapas têm mais conceitos intermediários como, por exemplo, os mapas de três termos base dos usuários 2, 4, 15 (APÊNDICE H) têm sete ou mais conceitos intermediários, enquanto os usuários 6 e 17 (APÊNDICE H) têm apenas 4 ou menos conceitos intermediários.

Além disso, na execução do protótipo, existe a possibilidade de configurar uma quantidade extra de conceitos intermediários. Assim, um mapa conceitual que já tenha conectado todos os termos do usuário pode ainda receber proposições adicionais. Um exemplo dessa diferença pode ser observado nos mapas com os termos de consulta ‘Jean Piaget’, ‘Educacional software’ e ‘Seymour Papert’, mostrados na subseção 4.5.7, Figura 74 e Figura 75. O primeiro mapa apenas faz a conexão dos termos de consulta, gerando 2 conceitos intermediários, e o segundo mapa contém, além dos elementos do primeiro mapa, três conceitos extras e seis proposições extras.

O Gráfico 17 relaciona o nível de aceitação do mapa, pelo usuário, com a proporção de novos conceitos sobre a quantidade total de conceitos no mapa. Como foram poucas avaliações realizadas (47 avaliações feitas por 17 usuários) não se pode ter uma conclusão estatística, mas, observa-se uma indicação de tendência de melhoria da aceitação do mapa com a diminuição da quantidade de novos conceitos. De fato, existe configuração no algoritmo para que os mapas resultantes sejam primordialmente simples e claros, ao invés de demasiadamente completos ou com todas as relações possíveis. Assim, nem todas as relações encontradas são apresentadas no mapa, pois “[...] num mapa conceitual existe sempre um compromisso entre ser claro e ser completo” (MOREIRA; MASINI, 1982, p. 49). Além disso,

resgatando a gestão da precisão citada por Araújo Junior (2007), na seção 2.4.5 do referencial teórico, um dos itens indicados é o controle da revocação e da exaustividade que têm como objetivo o aumento do índice de precisão. Assim, quanto mais itens recuperados, há tendência para menor precisão.

Gráfico 17 – Relação entre a quantidade de novos conceitos inseridos no mapa com o nível de aceitação do mapa conceitual



Fonte: Elaboração própria

Contudo, observa-se ainda que usuários podem desejar relacionamentos extras entre os conceitos. Um caminho para resolver essa demanda, diferente para cada necessidade informacional, é estender a interação do usuário no processo final de recuperação, dando-lhe oportunidade para escolher quais proposições extras ele deseja. Isso poderia ser feito com a utilização de *interactive information retrieval* (subseção 2.4.2.1), também sugerido nas conclusões, subseção 6.2.

5.5 Contexto dos experimentos sobre o modelo

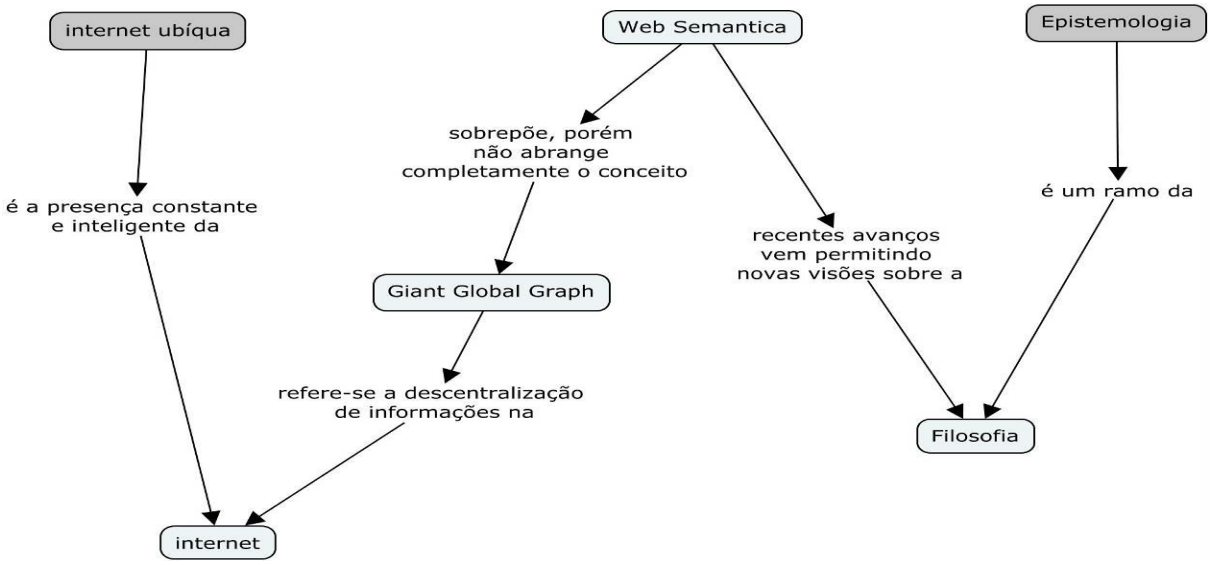
Essa seção discute o quanto são importantes as iterações e a retroalimentação no algoritmo do modelo, e sobre possíveis balanceamentos dos pesos atribuídos às métricas.

5.5.1 A importância das iterações e a retroalimentação

O principal ciclo de iterações do modelo corresponde ao processo de retroalimentação onde os nós da rede de informação, após ranqueamento e seleção, dão entrada novamente enquanto termos de consulta do usuário para serem consultados na base de conhecimento e, assim, expandirem a rede. Esse ciclo pode acontecer várias vezes e é representado pelo laço (4) do algoritmo do Quadro 7 na subseção 4.5.4.

Essas iterações acompanhadas da retroalimentação têm dois propósitos. O primeiro é possibilitar relações mais relevantes, pois, mesmo que todos os termos base já estejam unidos em único componente conectado, um número maior de iterações permite expandir a rede intermediária possibilitando a descoberta de relações através de ranqueamento em um universo maior, isto é, com maior possibilidade de revelar elementos relevantes.

Figura 82 – Mapa conceitual resultante dos termos “internet ubíqua” e “Epistemologia” após duas iterações



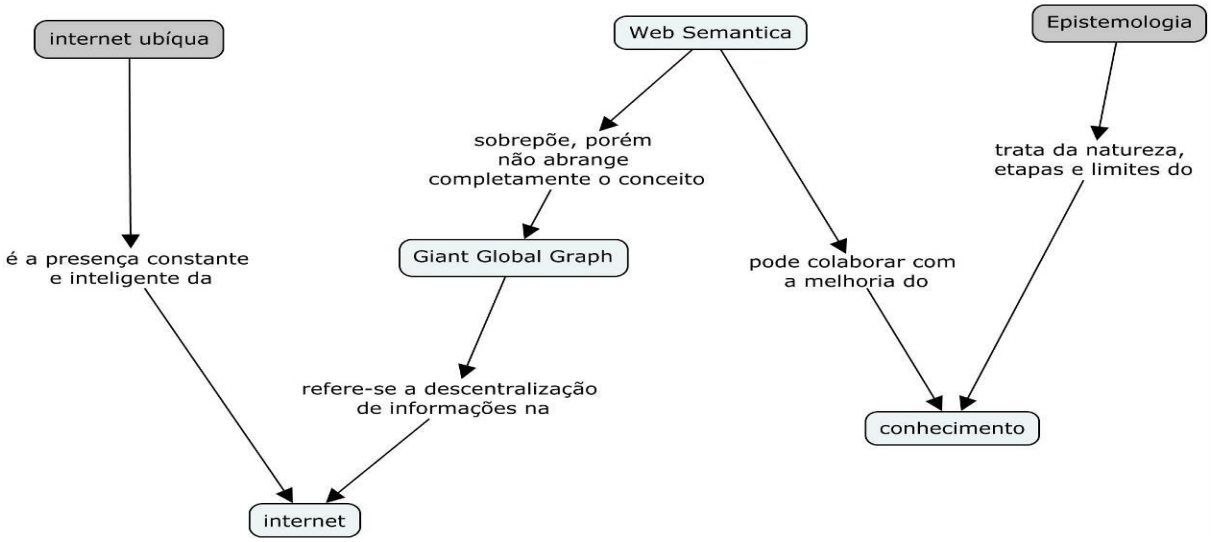
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

A Figura 82 representa um mapa conceitual resultante de um experimento sobre os termos ‘internet ubíqua’ e ‘Epistemologia’ na base de conhecimento privada apresentada na

subseção 4.5.6. Esse mapa é o resultado de um processo configurado para duas iterações e quatro conceitos intermediários.

Uma nova execução, porém, configurando o número de iterações para três, trouxe o mapa conceitual mostrado na Figura 83. Observa-se o conceito intermediário ‘Filosofia’, do mapa anterior (Figura 82), foi trocado por ‘conhecimento’. Devido à expansão da rede foi possível encontrar um conceito com um melhor índice da função de ranqueamento.

Figura 83 – Mapa conceitual resultante dos termos “internet ubíqua” e “Epistemologia” após três iterações

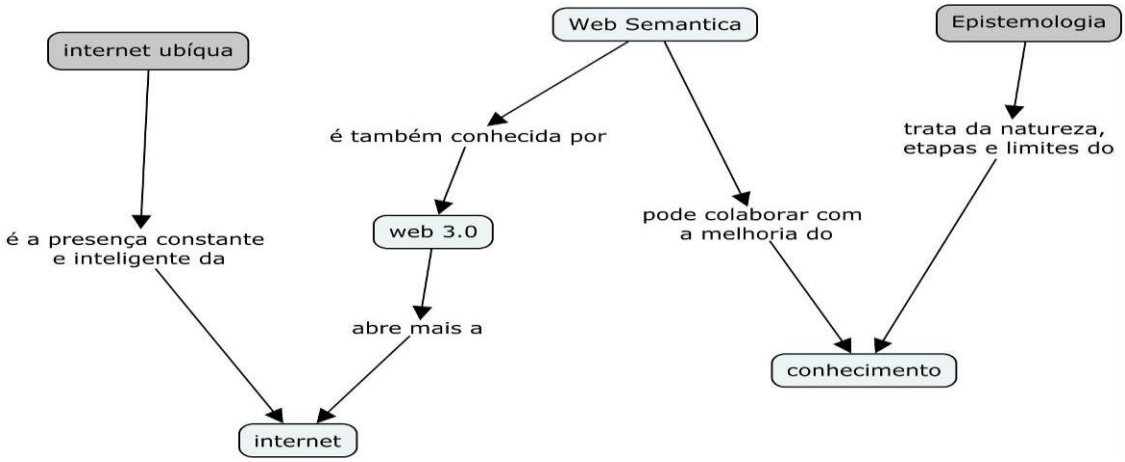


Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

A última execução, apenas modificando a quantidade de iterações de quatro para cinco, encontrou o mapa da Figura 84, onde o termo ‘Giant Global Graph’, dos mapas anteriores (Figura 82 e Figura 83), foi trocado por ‘web 3.0’, considerado mais importante no processo de ranqueamento.

Considerando a base de conhecimento usada e o contexto dos conceitos envolvidos ‘Web Semântica’ e ‘internet’, de fato, ‘web 3.0’ é mais relevante que ‘Giant Global Graph’. Da mesma forma, considerando os conceitos ‘Web Semântica’ e ‘Epistemologia’, de fato o conceito ‘conhecimento’ é mais relevante que ‘Filosofia’. Isto é, ‘Filosofia’, no geral, poderia ser mais importante e com mais conexões em outro contexto, diferente do contexto tratado nessa busca.

Figura 84 – Mapa conceitual resultante dos termos “internet ubíqua” e “Epistemologia” após 5 iterações



Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

O segundo propósito das iterações acompanhadas da retroalimentação é mais pragmático, ou seja, unir os termos base. Por exemplo, o mapa resultante do usuário 10 (disponível no APÊNDICE H, Figura 116) precisou de três iterações para unir os termos ‘Desejo’, ‘Arthur Schopenhauer’ e ‘Vegetarianismo’ e com uma rede intermediária de 20 mil nós. O usuário indicou na avaliação a inexistência da proposição ‘Desejo é um conceito central para Shopenhauer’. Contudo, na base de conhecimento não existia relação direta entre os dois conceitos, mas o algoritmo achou quatro conceitos intermediários e conseguiu estabelecer a relação, com um mapa conceitual de diâmetro igual a oito.

Seguem outros exemplos de convergência tardia: o mapa conceitual resultante dos termos do usuário 7 (disponível no APÊNDICE H, Figura 113) convergiu com 8 iterações, ligando os termos ‘Gamification na educação’, ‘Jogo’, ‘Digital media’, ‘Formação docente’, ‘Teorias de Ensino’ e ‘Autonomia’; o usuário 16 forneceu termos pertencentes a diversas áreas: ‘Cavaleiros do Zoodíaco’, ‘MasterChef’, ‘Séries de TV’, ‘Futebol’, ‘Copa do Mundo FIFA’ e ‘Olimpíadas 2016’ e o processamento somente convergiu para um único componente conectado depois de 15 iterações (mapa conceitual disponível no APÊNDICE H, Figura 127).

5.5.2 Pesos relativos às métricas de rede

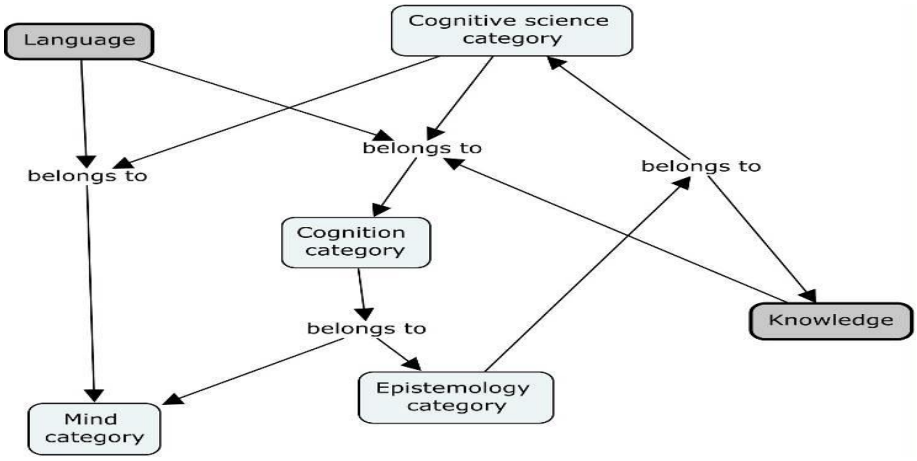
As métricas de rede são usadas no algoritmo do modelo, Quadro 7 da subseção 4.5.4, principalmente, para a ranqueamento que seleciona os nós que serão usados na retroalimentação em cada iteração do algoritmo. Nessa fase do algoritmo, laço 4 do Quadro 7

da subseção 4.5.4, existem dois *rankings* independentes para os nós da rede, o primeiro é uma composição de *betweenness* com *closeness,* e o segundo é pelo cálculo direto de *eigenvector*.

Observou-se ao longo dos testes piloto e validação com usuários, que os mapas tendiam para dois polos: (i) relacionamentos com conceitos intermediários categorizados como individuais, segundo a Teoria do Conceito de Dahlberg (1978), discutida na subseção 2.2.1 do referencial teórico, que representam instâncias diretas tais como nomes de pessoas, de universidades; ou (ii) relacionamentos com conceitos gerais que, segundo Dahlberg, seriam como classes que representam um grupo de conceitos individuais, como no caso de 'Universidade'.

O caso onde predominaram os conceitos individuais acontecia quando o peso das métricas *betweenness+closeness* era maior, e o caso com maior número de conceitos gerais ocorria quando *eigenvector* era maior. Esses dois polos podem ser observados nos mapas conceituais da Figura 85 e Figura 86. O peso maior para *eigenvector* fez aparecer conceitos mais gerais: categorias de algumas áreas do conhecimento. Por outro lado, o peso maior para *betweenness+closeness* revelaram conceitos individuais, como 'Jawdat Said' e 'Centre for Independent Social Research'.

Figura 85 – Mapa conceitual resultante dos termos 'Language', e 'Knowledge' com peso maior no fator *eigenvector*

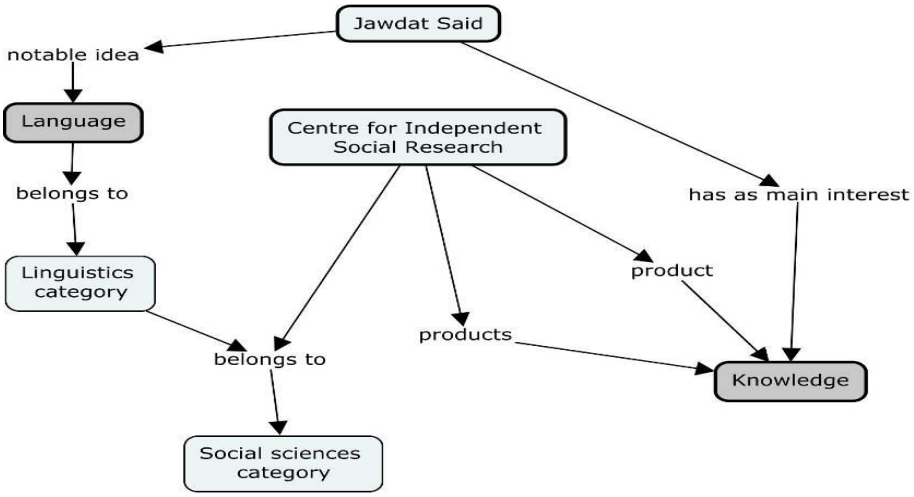


Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Uma possibilidade de explicação desse fenômeno é que a métrica *eigenvector* encontra *hubs* com mais facilidade, sendo esses conceitos mais gerais por estarem ligados a tantos outros nós, enquanto a métrica *betweenness+closeness* encontra elementos de intermediação que estão próximos ao centro, isto é, elementos que nem sempre possuem muitas conexões.

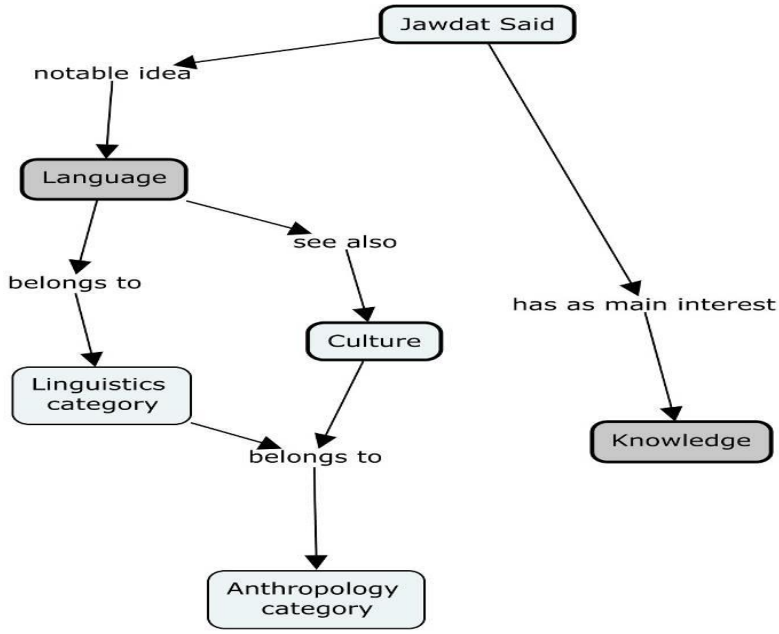
Essas diferenças topológicas podem ser observadas nos exemplos mostrados na Figura 29, Figura 30 e Figura 31 da seção 2.5.4 sobre medidas e métricas de rede, no referencial teórico.

Figura 86 – Mapa conceitual resultante dos termos ‘Language’, e ‘Knowledge’ com peso maior no fator *betweenness+closeness*



Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 87 – Mapa conceitual resultante dos termos ‘Language’, e ‘Knowledge’ com pesos balanceados entre os fatores *eigenvector* e *betweenness+closeness*



Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

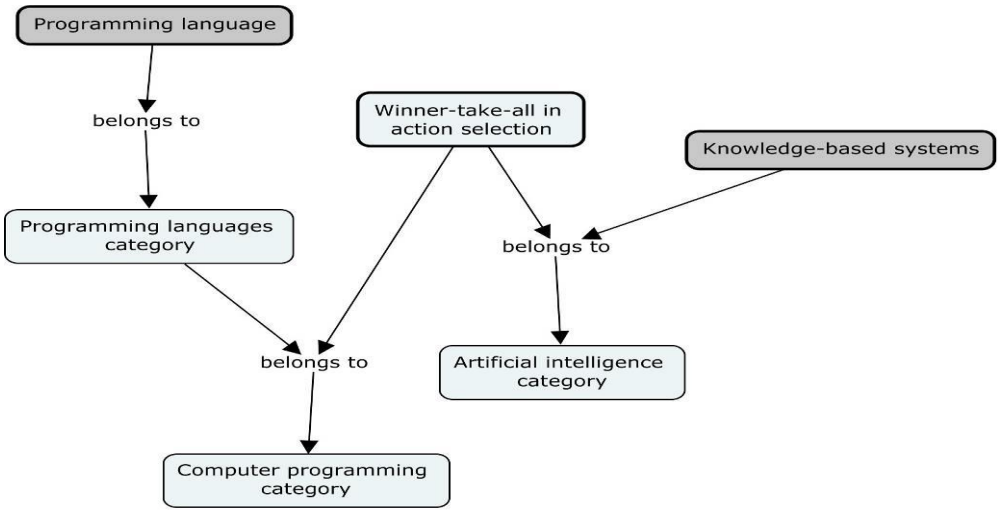
Uma solução que visa equilibrar a quantidade de relacionamentos entre conceitos gerais e individuais, é configuração do ranqueamento com o mesmo peso para as métricas

eigenvector e *betweenness+closeness*. O mapa conceitual da Figura 87, foi gerado com essa sugestão de parâmetros.

Porém, esse equilíbrio somente é possível, é claro, se houver conceitos intermediários individuais e gerais suficientemente adequados entre os termos base. Por exemplo, o mapa resultante dos termos base do usuário 1, APÊNDICE H - Figura 102, somente encontrou conceitos gerais. Dessa forma, o mapa se apresentou demasiadamente simples e hierárquico.

Uma solução mais adequada para uso do modelo proposto é fornecimento, por parte do usuário, de termos mais específicos. Assim, ao invés de 'Language', preferir algo mais específico, como 'Programming language', ou ainda, ao invés de 'Knowledge', preferir 'Knowledge-based system'. O resultado dessa especialização nos termos pode ser conferido no mapa da Figura 88.

Figura 88 – Mapa conceitual resultante dos termos 'Programming language', e 'Knowledge-based system' com pesos balanceados entre os fatores *eigenvector* e *betweenness*



Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Finalmente, a solução definitivamente adotada no modelo proposto foi a introdução de uma heurística para fazer esse balanceamento, como descrita passo 13 do algoritmo do Quadro 7 - subseção 4.5.4. Tal como já explicitado na subseção 4.5.5 que descreve a execução do protótipo, essa heurística não foi implementada. Nos testes piloto e na validação ela foi executada manualmente.

5.6 Contexto das avaliações dos usuários

Essa seção analisa e discute os resultados das avaliações dos usuários. Além disso, analisa a relação existente entre as métricas precisão e revocação.

5.6.1 Avaliações diretas dos usuários

Analisando de forma conjunta o Gráfico 8, Gráfico 9 e Gráfico 10 da subseção 4.6.2 referente às avaliações dos aspectos (i) **entendimento**: o quanto o mapa auxilia o entendimento das relações entre os termos base, (ii) **pesquisa**: o quanto o mapa auxilia como ponto de partida para uma pesquisa sobre as relações dos termos base, e (iii) **construção mapa**: o quanto o mapa auxilia como ponto de partida na construção de um mapa conceitual sobre as relações entre os termos base, e que também avalia, separadamente, os três mapas conceituais (3 termos usuário, 6 termos usuário, e 3 termos comum), observa-se que não houve diferença significativa entre o mapa com três termos base e outro com seis termos base, ambos fornecidos pelo usuário. Contudo, o mapa comum que todos avaliaram obteve um resultado um pouco melhor. Possivelmente por ter sido um mapa escolhido pelo autor dessa pesquisa, em função de sua clareza, dentre aqueles resultantes dos experimentos iniciais.

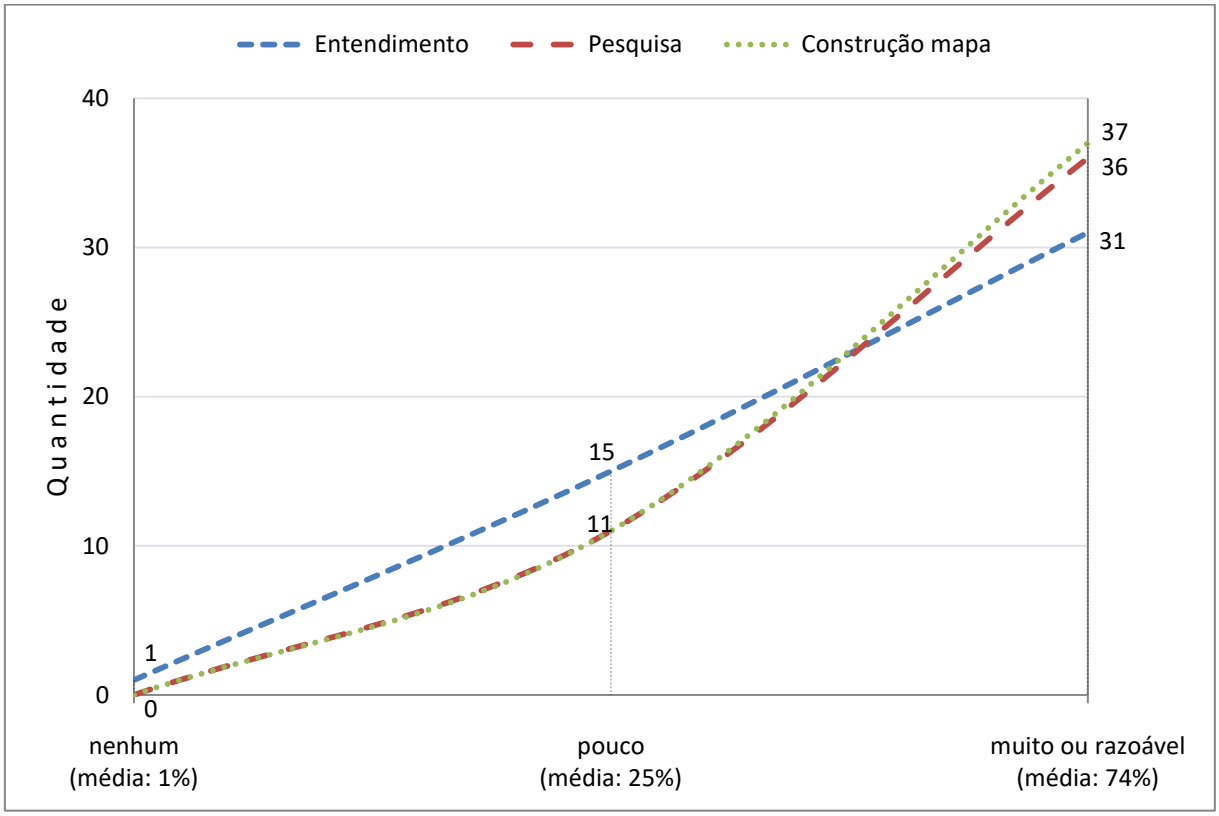
Consolidando essas avaliações, e numa interpretação qualitativa, observa-se que o resultado foi bom, pois houve 104 (37+36+31) indicações de razoável e muito bom, contra 37 (11+11+15) indicações de pouco, e apenas uma indicação de nenhum, como pode ser observado no Gráfico 18.

Observa-se que as avaliações foram melhores nos aspectos que usam a informação recuperada como ponto de partida para continuidade da tarefa cognitiva, ou seja, nos itens '**pesquisa**' e '**construção do mapa**', em comparação com aspecto que examina a informação recuperada como acabada, '**entendimento**'.

À luz da equação de Brookes, $K[S] + \Delta I = K[S+ \Delta S]$, discutida na subseção 2.1.4 do referencial teórico, há indicativos de que esses resultados possam ser interpretados como ponto de partida para a construção do conhecimento a partir da informação recuperada, ou mapa conceitual resultante ΔI. Quando o usuário recebe o mapa resultante, este se relaciona com o seu conhecimento prévio $K[S]$, podendo dar continuidade a sua pesquisa ou à construção do mapa a partir do novo conhecimento recém-criado $K[S+ \Delta S]$. A aprendizagem significativa de Ausubel, discutida na subseção 2.2.5 do referencial teórico, também estabelece que o usuário usa a nova informação recebida, nesse caso, o mapa conceitual

resultante, em conjunto com os seus subsunçores, ou conhecimento prévio, para modificar a sua estrutura cognitiva dando novos significados aos subsunçores, ocorrendo, assim, a diferenciação progressiva (subseção 2.2.5). Na medida em que o usuário continua o processo de pesquisa ou construção do mapa, ocorre a reconciliação integradora (subseção 2.2.5), onde os subsunçores que ganharam novos significados provocam modificações em outros subsunçores que estão conectados a eles trazendo novas modificações a sua estrutura cognitiva. Todo esse movimento acontece dentro do paradigma cognitivo da CI, descrito por Capurro (2003) na seção 2.1.2 do referencial teórico que foi proposto inicialmente por Brookes.

Gráfico 18 – Comparação entre as três avaliações: auxílio no entendimento, auxílio em pesquisa e auxílio na construção de mapa conceitual, considerando os mapas dos usuários e o mapa comum



Fonte: Elaboração própria

Dentre o universo de mapas conceituais avaliados, um caso típico que pode representar essa situação é a avaliação realizada pelo usuário 7 no mapa de três termos, Figura 112 do APÊNDICE H, sobre ‘Mundo líquido’, ‘Sociologia’ e ‘Rede social’. A avaliação do

mapa segundo o aspecto ‘entendimento’ não foi boa, porém, obteve uma avaliação muito boa no aspecto ‘pesquisa’ e melhor ainda no aspecto ‘construção do mapa’.

5.6.2 Avaliações pelas métricas de RI

As métricas para avaliação da qualidade em RI foram apresentadas na subseção 2.4.5 do referencial teórico e a metodologia para a sua aplicação no presente trabalho foi mostrada na subseção 3.5.4. Essa seção discute os resultados apresentados na subseção 4.6.3 tendo como base essas referências.

O Gráfico 11 e o Gráfico 12 da subseção 4.6.3 apresentam o resultado do cálculo da métrica ‘precisão’ no contexto dos novos conceitos e das proposições recuperadas, diferenciando os três mapas avaliados. Observa-se semelhança nos valores da ‘precisão’ entre os mapas produzidos a partir de três e seis termos do usuário. Porém, o mapa comum a todos os usuários obteve um resultado um pouco melhor, que pode ser atribuído ao fato dele ter sido escolhido pelo autor dessa pesquisa, em função de sua clareza, dentre aqueles resultantes do teste piloto.

Mesclando as informações dos dois gráficos citados, no Gráfico 19, observa-se que a maior parte dos novos conceitos e das proposições recuperadas foram relevantes para os usuários, com índice entre 60% a 100%.

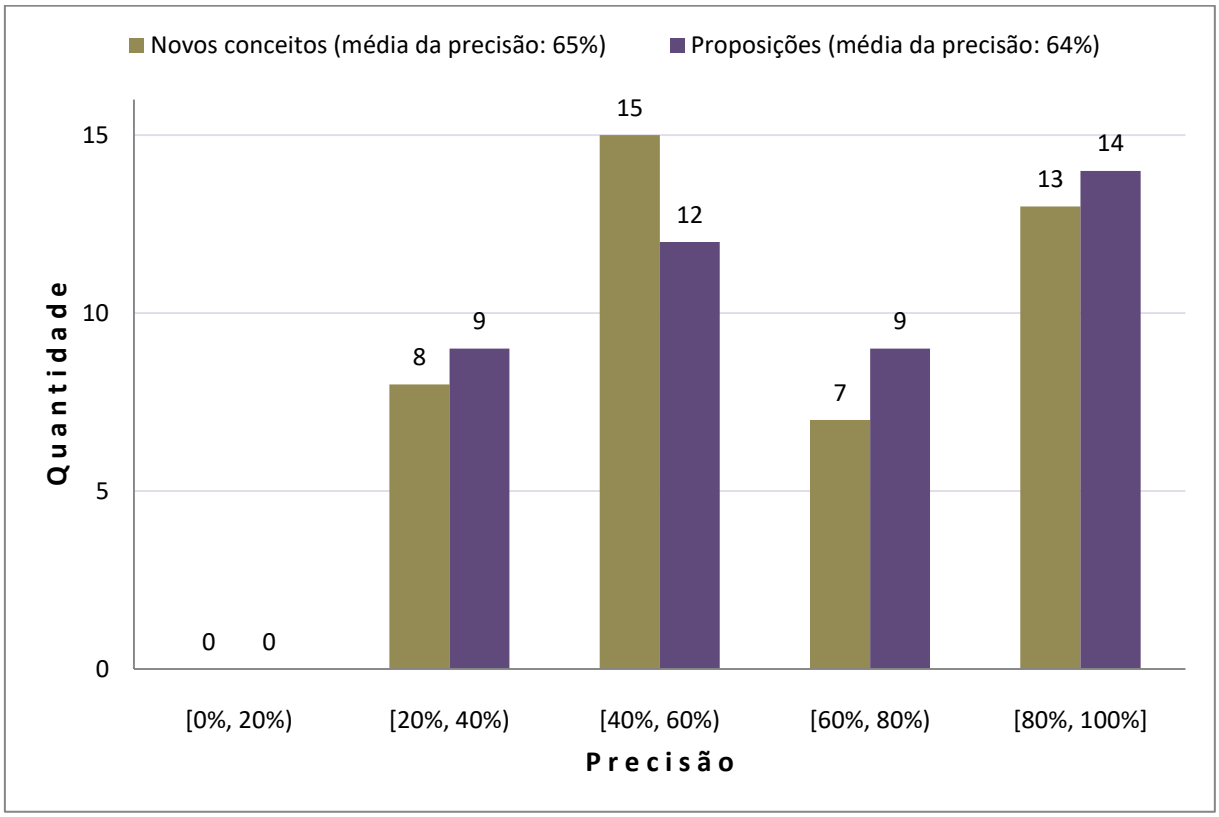
Quanto à revocação das proposições recuperadas, representadas pelo Gráfico 13 da subseção 4.6.3, os índices de recuperação das proposições foi muito bom, atingindo a média de 99% nos três mapas avaliados. Sendo que o mapa de três termos do usuários houve revocação máxima, de 100%.

As proposições indicadas pelos usuários como faltantes no mapa, só foram computadas se existissem na base de conhecimento. O método para verificação de sua existência está explicado no capítulo da metodologia na subseção 3.5.5.1, na parte denominada de ‘Métricas para avaliação da qualidade das proposições recuperadas’.

Dessa forma, do total de 139 proposições indicadas, a grande maioria, 135 proposições, não pertenciam a base de conhecimento DBpedia. Por exemplo, o usuário 10, APÊNDICE H e Figura 116, assinalou a falta das proposições ‘Shopenhauer é conhecido como filósofo da Angústia e Sofrimento’ e ‘Desejo representa Angústia e Sofrimento’. Contudo essas proposições não existem na DBpedia. Em outro caso, o usuário 12, APÊNDICE H e Figura 118, indicou que poderia haver um relacionamento mais

direto entre ‘Artista’ e ‘Surrealismo”, passando por ‘Sigmund Freud’, ‘André Breton’ ou ‘Salvador Dalí’. Porém, nenhuma dessas relações existem na DBpedia.

Gráfico 19 – Precisão da informação recuperada



Fonte: Elaboração própria

Uma consequência negativa desse fato, é que o usuário pode desqualificar o mapa conceitual como um todo. Isso porque o participante dessa validação conhece bem o assunto tratado no mapa, conforme pré-requisitos discutidos na subseção 3.5.2. Essa característica faz o usuário aumentar a expectativa quanto a presença de relacionamentos que, para ele, são óbvios ou fundamentais, considerando o seu alto nível de exigência enquanto conhecedor daquele assunto.

Observou-se que alguns usuários não entenderam que a indicação das proposições faltantes deveriam ter o objetivo de relacionar os termos base, seja diretamente ou indiretamente por meio de outros conceitos. Por exemplo, o usuário 6, APÊNDICE H, Figura 110 e Figura 111, identificou a falta de um relacionamento do termo ‘Robótica educacional’ com ‘Lego Logo’. E no outro mapa ele identificou a falta da proposição ‘Hápkido pertence à Categoria das Artes Coreanas’. Porém, tais proposições não estabelecem conexão entre os termos base, direta ou indiretamente. Outras situações ocorreram no mapa comum a todos,

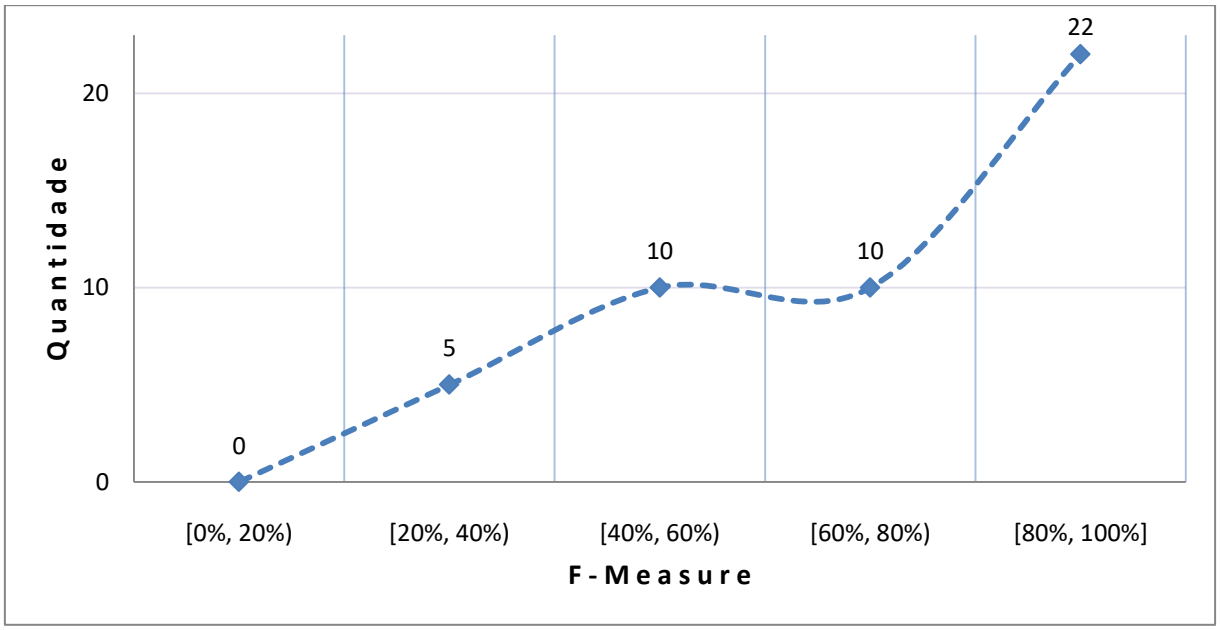
mostrado na Figura 37: o usuário 10 sinalizou a falta da proposição ‘Jean Piaget se relaciona com as operações de Cooperação e Coação’; o usuário 12 e sugeriu as proposições ‘Software educacional podem ser Simuladores, Jogos, Tutoriais’. Observa-se que, nesses casos, os usuários estavam querendo explicar melhor um conceito, mas não estabelecer relações entre os termos base.

Possivelmente houve falha no processo de comunicação com o usuário. Acredita-se que esse problema poderia ser minimizado se houvesse uma explicação melhor quanto à determinação das proposições faltantes.

Houve quatro casos de indicação da falta de uma proposição que era existente na base de conhecimento. Como exemplo, o usuário 10 sugeriu: ‘Construtivismo é uma proposta filosófica de Jean Piaget’ e foi encontrada uma tripla RDF semelhante: ‘Jean Piaget é conhecido pelo Construtivismo’.

Quanto a revocação dos novos conceitos recuperados, a subseção 3.5.5.2 justifica o porquê dele não ter sido calculada.

Gráfico 20 – F-measure das proposições recuperadas



Fonte: Elaboração própria

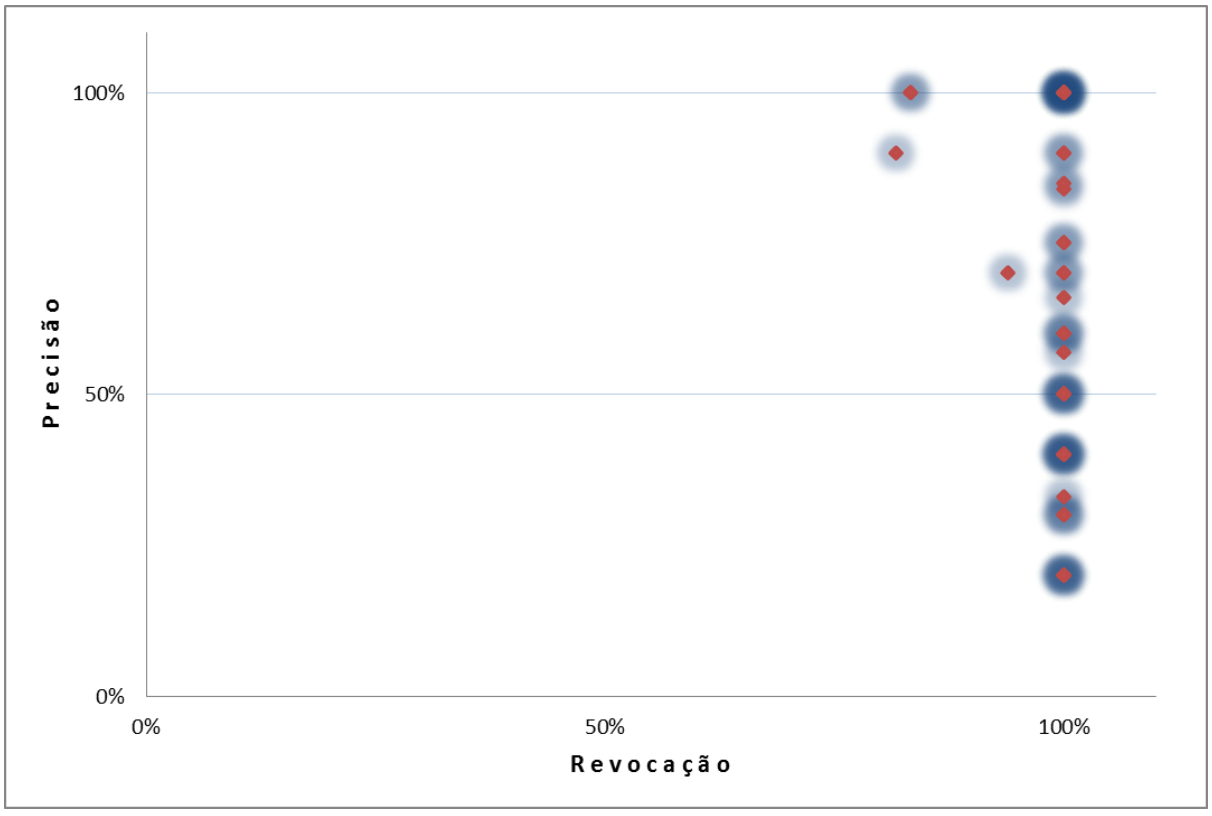
Quanto à métrica ‘F-measure’, abordada na subseção 2.4.5.2 do referencial teórico, o Gráfico 20 sintetiza a precisão e a revocação das proposições recuperadas com a apresentação da métrica ‘F-measure das proposições recuperadas’. Isto é, conforme a subseção 3.5.5.1 da metodologia, ela é a média harmônica entre a precisão e a revocação das proposições

recuperadas. Nessa síntese mostrada no Gráfico 20, observa-se um resultado geral e qualitativamente positivo do modelo de recuperação de informação e conhecimento, com a maioria das avaliações enquadradas na faixa de 60% a 100% de F-measure. Contudo, pelo baixo número de avaliações realizadas, 47 ao todo, esse resultado não é quantitativo e nem conclusivo.

5.6.3 Relação entre precisão e revocação

Tal como discutido na subseção 2.4.5.2, espera-se de uma recuperação de informação um relacionamento entre ‘precisão’ e ‘revocação’ como aquele mostrado no Gráfico 2. Contudo, observou-se um resultado diferente nas 47 medições de ‘precisão’ e ‘revocação’ em relação às proposições recuperadas, mostrado no Gráfico 21, e com alguma semelhança com o Gráfico 3, que mostra uma situação hipotética de uma recuperação de informação perfeita.

Gráfico 21 – Relação entre precisão e revocação das proposições recuperadas



Fonte: Elaboração própria

Analisando os trabalhos de Buckland *et al.* (1992) e Buckland e Gey (1994), tal como apresentados na subseção 2.4.5.2 do referencial teórico, onde eles afirmam que uma

recuperação em múltiplos estágios pode trazer um resultado com melhor nível de ‘precisão’ simultaneamente com um melhor nível de ‘revocação’, é possível achar algumas semelhanças do processo adotado por Bucklan *et al.*, com o processo do modelo apresentado nessa tese. As primeiras recuperações de informação na base de dados ligados trazem todas as triplas RDFs possíveis para a formação da rede de informação expandida, maximizando a revocação, tal como sugere Bukland *et al.* Na próxima etapa, os autores fazem outra busca no conjunto de informações recuperadas para melhorar a precisão. De forma análoga, no modelo do presente trabalho, a rede de informação é reduzida até a formação do mapa conceitual resultante, pela eliminação dos nós que não são selecionados pelos algoritmos de ranqueamento baseados nas operações de redes complexas e, desta forma, se aproximando de resultados com maior ‘precisão’.

5.7 Contexto da base de conhecimento

Essa seção analisa e discute possível relacionamento entre o tamanho da base de conhecimento com o resultado das avaliações dos usuários. Também faz comparações entre as duas bases de conhecimento usadas nos testes.

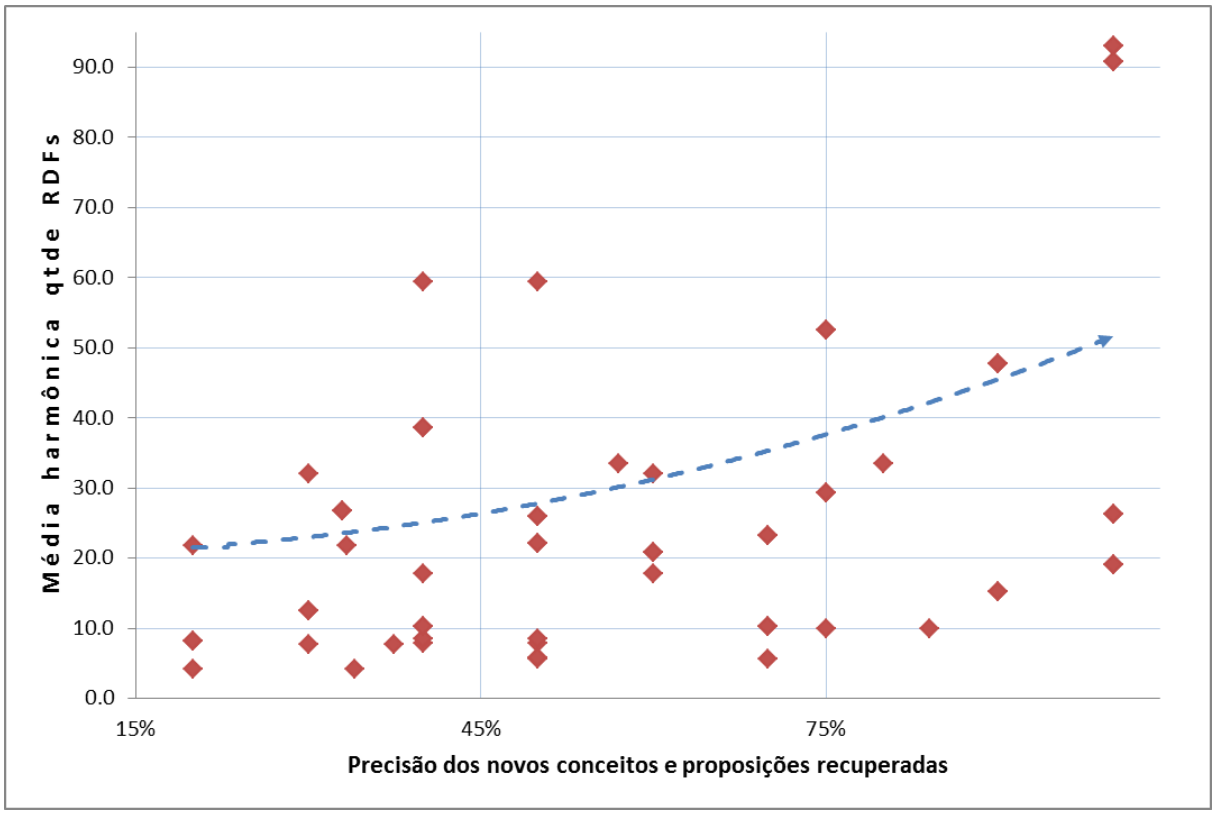
5.7.1 Quantidade disponível de RDFs na base de conhecimento

Apesar da base de conhecimento DBpedia possuir mais de 6 milhões de entidades, tal como descrito na subseção 2.3.5 do referencial teórico, e mais de 4,5 milhões na língua inglesa, como descrito na subseção 3.5.1 da metodologia, faltam ainda importantes entidades do mundo real bem como muitas relações entre sujeitos/recursos e objetos/valores. Isso foi constatado nos vários testes de recuperação de informação realizados sobre ela durante o desenvolvimento desse trabalho, sendo alguns casos identificados pelos usuários e relatados na subseção 5.6.2.

O Gráfico 22 relaciona precisão dos novos conceitos e proposições com a quantidade de triplas RDFs de cada termo base, sintetizados como a média harmônica entre eles. A curva do gráfico mostra uma tendência de quanto maior a quantidade de triplas RDFs associadas ao mapa, melhor é a avaliação de sua precisão. É claro que, devido às poucas avaliações existentes (47 avaliações feitas por 17 usuários), não é possível extrair uma conclusão estatística desse gráfico, mas, é interessante observar essa tendência, pois isso leva à reflexão

de que é preciso aumentar as bases de dados ligados na Web Semântica, tal como já sinalizado por diversos autores ao longo desse trabalho, como Berners-Lee (2010), Heath e Bizer (2011), Stuckenschmidt, Noessner e Fallahi (2012), Stuckenschmidt (2012), Auer *et al.* (2013), e também aumentar a quantidade de dados abertos como alertado por Berners-Lee (2010), Nazario, Silva e Rover (2012), Ding, Peristeras, Hausenblas (2012), Bauer, Kaltenböck (2012), Pedroso, Tanaka e Cappelli (2013).

Gráfico 22 – Relação entre a média harmônica das quantidades de RDFs dos termos base com a precisão dos conceitos novos e proposições recuperadas



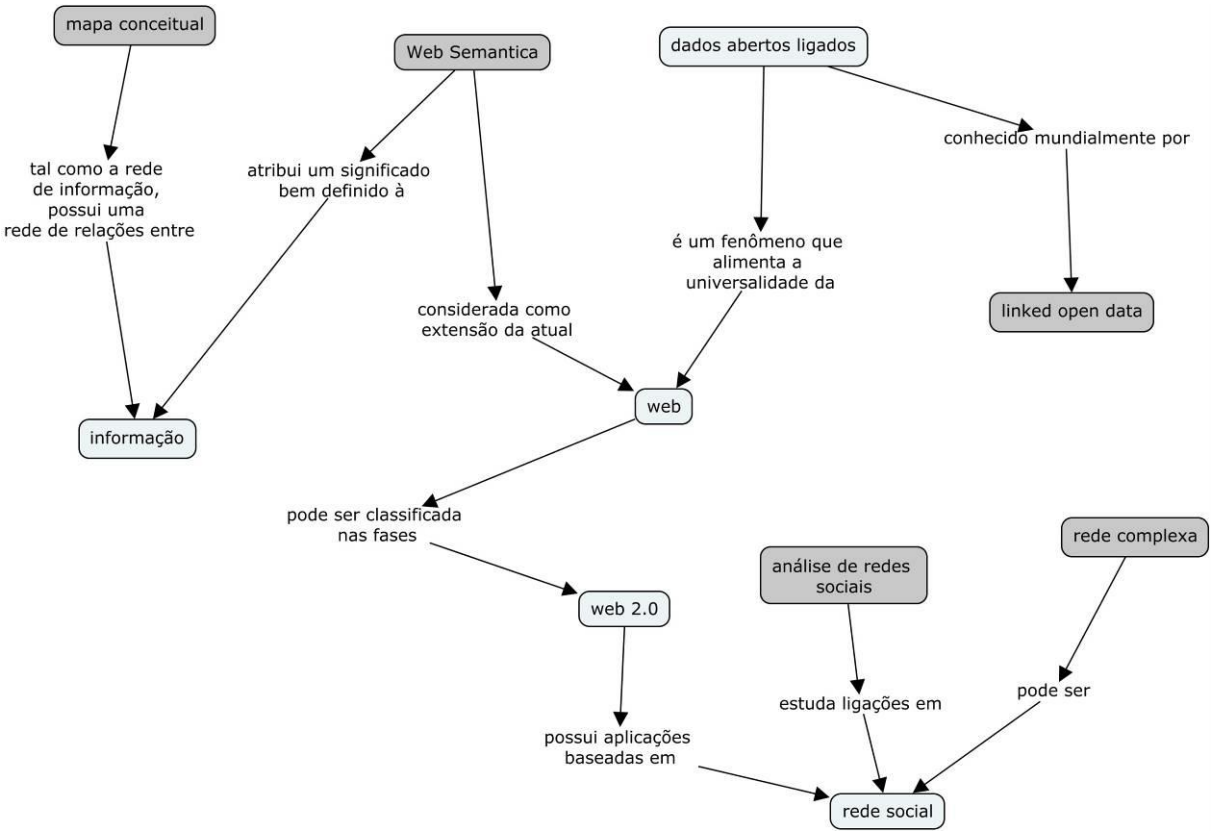
Fonte: Elaboração própria

5.7.2 Comparação das duas bases de conhecimento

As seções 4.5.6 e 4.5.7 apresentaram testes piloto sobre duas bases de conhecimento distintas: uma base privada e a DBpedia. Sendo essa última usada na validação do modelo. Essa seção faz uma breve comparação prática entre essas duas bases. Para essa comparação, foram selecionados os mesmos termos base para serem processados nas duas bases. São somente cinco termos e são associados a temas tratados no referencial teórico desse trabalho: ‘mapa conceitual’, ‘Web semântica’, ‘linked open data’, ‘análise de redes sociais’ e ‘redes

complexas'. Pelo fato da base selecionada na DBpedia para essa pesquisa estar em inglês, conforme já discutido na seção 3.5.1, os termos foram traduzidos para um conjunto equivalente: 'Concept map', 'Semantic web', 'linked open data', 'Social network analysis' e 'complex network'. O primeiro mapa conceitual, Figura 89, é resultante do experimento feito na base de conhecimento privada, e o segundo mapa, Figura 90, é relativo à base DBpedia.

Figura 89 – Mapa conceitual resultante de processamento na base de conhecimento privada

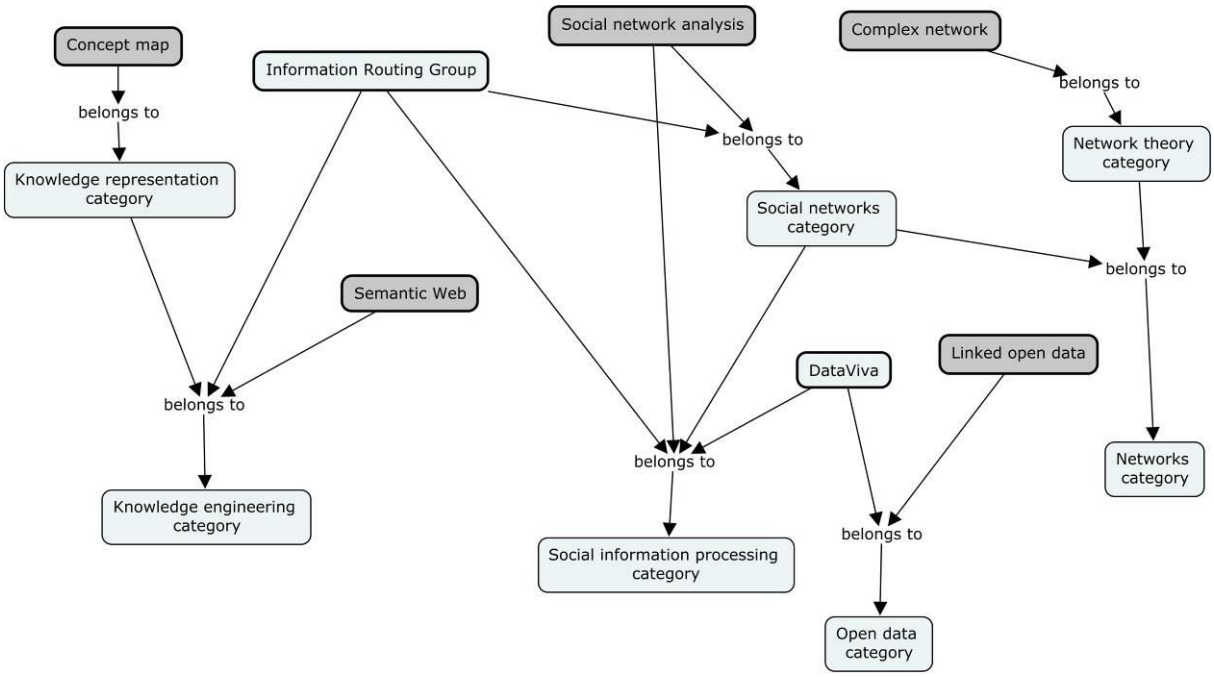


Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Numa avaliação mais geral, e possível observar que o primeiro mapa tem proposições mais interessantes do que o segundo mapa que tem, predominantemente, a frase de ligação 'belongs to' tornando-o mais hierárquico. Numa avaliação mais específica, observa-se que, pelo fato do primeiro mapa não se utilizar de ontologias ou quaisquer sistemas classificatórios sobre as frases de ligação, ele tem proposições bem descritivas e didáticas para o entendimento do leitor, porém, sem padronização. Já o segundo mapa se utiliza das ontologias associadas à DBpedia, como abordado na subseção 2.3.5 do referencial teórico. Assim o primeiro mapa, apesar de mais legível e provavelmente mais útil ao usuário, ele tem pouca interoperabilidade. Essa observação estende-se também à base de conhecimento privada como um todo, isto é, ela possui baixa interoperabilidade com outras bases de conhecimento. Um

encaminhamento de solução para esse problema seria a aplicação de ontologias, pois elas são usadas para integrar bases heterogêneos, permitindo a interoperabilidade entre sistemas diferentes (GRUBER, 2009). A subseção 2.3.4 do referencial teórico discute as ontologias para dados ligados. Outra reflexão a ser considerada é que se o número de triplas RDFs na DBpedia fosse maior, possivelmente o mapa conceitual resultante seria melhor. A necessidade de aumentar as bases de dados ligados já foi discutida na subseção 5.7.1 e é defendida por diversos autores.

Figura 90 – Mapa conceitual resultante de processamento na base de conhecimento DBpedia



Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

5.8 Análise comparativa com os trabalhos correlatos

Os trabalhos correlatos foram apresentados no capítulo da introdução na subseção 1.4 e essa subseção discute uma síntese comparativa desses trabalhos correlatos com o presente trabalho, conforme mostra o Quadro 13. A primeira coluna desse quadro contém a referência bibliográfica do trabalho e as outras colunas de 1 a 13, representam os critérios usados na comparação:

(1): recupera ou extrai informações;

(2): enquanto método de RI tem a recuperação de texto como forma predominante.

(3): aceita como ponto de partida uma lista de termos textuais fornecidos pelo usuário;

(4): usa dados ligados da web semântica como base de conhecimento;

(5): tem como foco a descoberta de relacionamentos existentes entre os termos fornecidos pelo usuário;

(6): revela relacionamentos intermediários entre os termos da busca ainda que estes estejam distantes por alguns nós e ligações ao longo da rede informacional;

(7): usa métrica de rede e algoritmos de grafos sobre a rede informacional como parte fundamental para o ranqueamento e seleção de documentos relevantes;

(8): considera a topologia da rede informacional para o ranqueamento dos documentos;

(9): usa um processo de retroalimentação onde, a partir de uma única solicitação do usuário, realiza novas buscas na base de conhecimento a partir de documentos já recuperados;

(10): usa um formato visual de rede informacional para apresentar a informação resultante;

(11): usa mapa conceitual para representar a informação resultante;

(12): parte da informação recuperada é apresentada ao usuário como conceitos de um mapa conceitual gerado de forma semiautomática;

(13): parte da informação recuperada é apresentada ao usuário como relações entre conceitos de um mapa conceitual gerado de forma semiautomática.

Quadro 13 – Comparação entre a proposta e os trabalhos correlatos

Referência para o trabalho	Critérios comparativos												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Proposta do presente trabalho	X	X	X	X	X	X	X	X	X	X	X	X	X
Cañas *et al.* (2004)	X	X	X							X	X	X	
Lima (2005)	X	X								X	X	X	
Thammasut e Sornil (2006)	X	X					X	X					
Graudina e Grundspenkis (2008)	X									X	X	X	X
Truong *et al.* (2008)	X	X					X	X					
Heim, Ertl e Ziegler (2010)	X	X	X	X	X								
Lohmann *et al.* (2010)	X	X	X	X	X	X			X	X			
Guéret *et al.* (2012)	X			X			X	X					
Valerio, Leake e Cañas (2012)	X	X								X	X	X	X
Paulheim (2013)	X	X	X	X									
McLinden (2013)	X						X	X		X	X	X	X
Cury, Perin, Santos Junior (2014)	X		X							X	X	X	X
Usbeck (2014)	X	X	X	X			X						

Fonte: Elaboração própria

O trabalho que mais se aproximou da presente proposta é o de Lohmann et al. (2010), com oito indicações de similaridade das treze totais. A proposta dos autores consegue, a partir de uma lista de termos do usuário, descobrir relacionamentos entre eles advindos de uma base de dados ligados, apresentando-os num formato de rede informacional próximo a um grafo de RDFs. Existe uma interface interativa adequada que permite grande flexibilidade nas consultas. Além disso, existe uma retroalimentação que consegue fazer a rede crescer na medida em que novos relacionamentos vão sendo descobertos. Contudo, após teste da ferramenta, disponível em http://www.visualdataweb.org/relfinder.php, observou-se que a descoberta de relacionamentos é mais indicada para aqueles com conexão direta, pois, vários testes realizados entre termos que se conectavam apenas indiretamente falharam. Adicionalmente, o trabalho não atende aos critérios de empregar operações de redes complexas na função de ranqueamento e nem mapas conceituais para apresentação de resultados.

A originalidade do presente trabalho concentra-se, principalmente, no conjunto de elementos de diversas áreas do conhecimento para a construção do modelo de RI em dados ligados, tais como o uso de conhecimentos em redes complexas e mapas conceituais. Os pontos específicos que mais se destacam são o uso de métricas de rede, algoritmos de grafos e análise topológica sobre a rede informacional como parte fundamental para o ranqueamento e seleção de nós mais relevantes, e a apresentação da informação recuperada no formato de um mapa conceitual.

6 CONCLUSÕES

O desenvolvimento do modelo para recuperação de informação e conhecimento usando operações de manipulação de redes complexas e geração de mapas conceituais, com foco no descobrimento de relacionamentos entre termos de uma consulta associada a uma necessidade informacional do usuário, cumpriu o objetivo geral do trabalho. O cumprimento dos objetivos específicos é discutido nos parágrafos seguintes.

À luz da equação fundamental da CI de Brookes e fundamentado nos paradigmas da CI, físico, cognitivo e social, o trabalho desenvolvido sugeriu um modelo para recuperação de informação e conhecimento. Sendo que, um dos maiores desafios no seu desenvolvimento foi a integração conceitual de várias áreas do conhecimento. Todavia, a natureza interdisciplinar da Ciência da Informação colaborou na tarefa de extrair elementos em cada área: Recuperação de Informação e Conhecimento, Visualização de Informação e Conhecimento, Web Semântica e Dados Ligados, Ciência das Redes, Mapas Conceituais e Aprendizagem Significativa. Um trabalho interdisciplinar, normalmente, possui desafios que, para serem vencidos, necessitam de um amplo e profundo conhecimento das áreas de sua atuação. Esse, talvez, seja o maior empecilho em trabalhos dessa natureza. Por outro lado, os resultados, normalmente, são interessantes do ponto de vista social e prático, uma vez que o mundo e seus problemas são predominantemente interdisciplinares. Com essa integração conceitual entre distintas áreas do conhecimento, cumpriu-se o primeiro objetivo específico elencado na introdução, na subseção 1.2.2.

O trabalho desenvolvido abriu possibilidades concretas de uso de elementos da Ciência das Redes, especificamente da análise de redes complexas, para a recuperação de informação e conhecimento dentro do contexto dos dados abertos ligados na Web Semântica. Destaque para o uso de operações em redes complexas no ranqueamento e seleção da informação recuperada. Também houve contribuições pragmáticas para a Ciência das Redes com a verificação de diversos fenômenos nas redes de informação recuperadas a partir de base de dados ligados, abrindo espaço para novas investigações.

O uso de mapas conceituais na apresentação dos resultados foi um ponto de grande relevância no trabalho. A aplicação das áreas de visualização de informação e conhecimento (*information visualization* e *knowledge visualization*) na recuperação de informação foi ao encontro dessa utilização de mapas conceituais no modelo proposto. A originalidade do trabalho aparece principalmente no uso de redes complexas e mapas conceituais na

recuperação de informação e conhecimento, conforme foi discutido na análise comparativa de trabalhos correlatos, subseção 5.8.

Destaca-se a busca de relacionamentos existentes entre os termos fornecidos pelo usuário, e não pela via mais tradicional com buscas de propriedades, definições ou explicações individuais de cada termo. Essas conexões entre os termos, ainda que por conceitos intermediários, vem ao encontro do que se discute na aprendizagem significativa no contexto dos mapas conceituais e na equação fundamental da CI de Brookes. A busca por relacionamentos abre um espectro de possibilidades interessantes em várias áreas do conhecimento e na disponibilização de serviços pra a sociedade. Por exemplo, cidadãos podem usufruir de um serviço com essas características para encontrar relacionamentos em informações governamentais de forma a terem uma postura mais ativa quanto ao acompanhamento de dados num contexto de transparência governamental e combate a corrupção.

O método exploratório e cíclico, de desenvolvimento do modelo e do protótipo foi determinante para a descoberta de parâmetros de cálculos e transformações de análise de redes que melhorassem empiricamente a síntese do mapa conceitual resultante. A automatização de boa parte do modelo, por intermédio do protótipo desenvolvido, foi fundamental para a descoberta de melhores caminhos na concepção do algoritmo bem como na realização de uma quantidade razoável de testes e na execução da validação. Com a realização dessa investigação cumpriu-se o segundo e terceiro objetivos específicos elencados na introdução, subseção 1.2.2.

Apesar da amostra de usuários que participou da validação, ser pequena, com 17 pessoas e quantidade total de 47 avaliações, existiram indícios qualitativos de que o modelo proposto produziu uma boa recuperação de informação. As melhores avaliações aconteceram sob a ótica do mapa conceitual resultante enquanto ponto de partida para a uma pesquisa mais profunda sobre as relações entre os termos fornecidos pelos usuários, ou como ponto de partida para a construção de um mapa conceitual mais completo. Além disso, o nível de precisão das proposições e novos conceitos recuperados foi razoável, sendo que a revocação das proposições recuperadas atingiu um índice de quase 100%. Um fator que deve ser levado em consideração é que os usuários eram especialistas dos assuntos abordados pelos mapas. Acredita-se que essa situação os deixou mais exigentes ou criteriosos no momento da avaliação. Se fossem usuários com uma necessidade informacional real, as informações recuperadas poderiam atender melhor as suas expectativas. Contudo, tal como previsto e discutido na seção 3.5.2 da metodologia, a exigência de conhecimento prévio sobre os termos

pesquisados seria importante para garantir uma correta avaliação, principalmente na indicação de proposições que faltaram. Com essa validação sobre o protótipo desenvolvido, cumpriu-se o quarto objetivo específico elencado na introdução, subseção 1.2.2.

O uso de tecnologias específicas da área de acesso à Web Semântica, manipulação de redes de informação e construção de mapas conceituais permitiu a descoberta de caminhos interessantes para a integração dos elementos e criação do protótipo. Se por um lado o protótipo deveria ser desenvolvido num tempo curto, dada a natureza dessa pesquisa, por outro, ele teria que ser eficiente, isto é, com boa velocidade de processamento para que fosse possível a execução dos testes e da validação do modelo. Ter esses dois elementos juntos é difícil, pois alta produtividade, normalmente, requer o uso de ferramentas prontas que, tradicionalmente, são mais lentas. Em busca do meio termo foram escolhidas bibliotecas para manipulação de redes que oferecessem os cálculos de métricas prontos e que, ao mesmo tempo, pudessem ser integradas às ferramentas de visualização para a inspeção visual como parte da metodologia. Essa trajetória de escolha não foi simples, pois bibliotecas para redes bem organizadas e com um bom aparato de comunicação com outros módulos eram lentas, e outras com alta performance, porém, complexas e com baixo nível de reutilização. A solução aconteceu em torno da utilização de partes de uma ou de outra de acordo com as suas peculiaridades que satisfizessem as necessidades do protótipo que estava sendo desenvolvido. Essas anotações e documentação estão disponíveis no código fonte do protótipo e podem dar um ganho de produtividade a futuros desenvolvedores nessas áreas trabalhadas.

A demanda pela ampliação das bases de conhecimento de dados abertos ligados é notória e recomendada por vários autores, como discutido ao longo desse trabalho. Esse aumento possibilita a disponibilização de serviços que atendam de forma diferenciada as demandas da sociedade. Uma possibilidade que se discute muito é a adoção de *crowdsourcing*. Outra forma, indireta, é aumentar o desenvolvimento de projetos para uso de dados abertos ligados, pois eles estimulam a organização de movimentos com o objetivo de aumentar as bases de conhecimento, retroalimentando, assim, a possibilidade para novos projetos. Portanto, projetos que envolvam o uso de dados abertos ligados, além de, tradicionalmente, oferecem um serviço diferenciado para os usuários permitindo ações que antes não eram possíveis, também estimulam movimentos para padronização, correção e ampliação, tanto de dados abertos quanto de dados ligados.

6.1 Limitações

A não automatização completa do modelo limitou a sua validação a um conjunto pequeno de usuários. A primeira parte do modelo, que consiste na normatização dos termos de consulta do usuário é feita manualmente, pois não há interface para esse propósito. Outra parte que ocupa um tempo razoável no processo, devido também à intermediação humana, é a execução da heurística que faz o balanceamento do mapa resultante com uma quantidade equilibrada de conceitos individuais e gerais. Isso requer uma nova execução com a modificação dos pesos das métricas de rede usadas. Além disso, a *layout* final dos conceitos no mapa conceitual precisa de auxílio humano para torná-lo mais legível.

A validação ficou limitada a uma quantidade reduzida de 17 usuários participantes. Ao todo foram 32 mapas conceituais avaliados e 47 avaliações realizadas ao todo. Apesar dos resultados qualitativos, essa amostra não permitiu a obtenção de resultados quantitativos estatisticamente validados.

6.2 Sugestões para novas pesquisas

A adoção de uma RI interativa (*interactive information retrieval*) no modelo poderia oferecer ao usuário maior flexibilidade em vários aspectos. São sugeridas três situações para a interação do sistema de RI com o usuário. Porém, apesar da existência de indícios de que a mudança de abordagem de RI para RI interativa seja benéfica para a melhoria do atendimento da necessidade informacional do usuário, é necessário investigar cada caso.

(i) Na escolha final dos conceitos, isto é, ao invés dele receber um mapa conceitual pronto, o usuário participaria de forma interativa do processo de ranqueamento e seleção dos conceitos que seriam mostrados no mapa final. Se essas escolhas forem realizadas ainda no meio do processo, seria possível fazer o algoritmo da RI percorrer um trajeto diferente dentro do universo dos dados abertos ligados e, assim, oferecer um mapa conceitual resultante também diferente.

(ii) Na determinação inicial da configuração relativa à escolha de elementos para modificar os processos de ranqueamento e seleção de nós das redes intermediárias. Essa configuração poderia, interativamente, ser modificada ao longo do processo e, assim, o usuário poderia avançar ou retroceder sobre algum ou outro valor de configuração até chegar em um resultado mais próximo de seu desejo informacional.

(iii) Na visualização do mapa conceitual resultante em formato tridimensional, com conceitos e frases de ligação mais relevantes em um plano mais próximo do usuário e, proporcionalmente, proposições menos relevantes, mais afastadas do usuário. De forma interativa, o usuário poderia trazer para um plano mais próximo dele os conceitos que seriam, eventualmente, mais relevantes a partir de rápido julgamento feito no mesmo momento de sua interação. Essa interface possibilitaria uma recuperação de informação mais ampla, isto é, com mais proposições que relacionassem indiretamente os termos de consulta do usuário. Porém, sem congestionar com excesso de conceitos o mapa conceitual mais próximo do usuário, pois os demais estariam em planos mais afastados e, devido a perspectiva, com tamanho reduzido. Além disso, conceitos e frases de ligação julgados, naquele momento, como não relevantes poderiam ser facilmente descartados pelo usuário.

A realização de testes em bases de conhecimento diferentes, como as *Linked Open Government Data* (LOGD), bases brasileiras tal como o Portal da Transparência do Governo do Brasil entre outras são ações que precisam ser experimentadas. Também o estudo mais aprofundado de ontologias para uma possível criação de um vocabulário adequado e baseado nos elementos de relacionamento encontrados nas triplas RDFs ou a adoção de alguma padronização para viabilizar o acesso a outras bases de conhecimento. Por exemplo, o padrão universal proposto pelo *Simple Knowledge Organization System*[79] (SKOS). Todas essas possibilidades precisam ser pesquisadas.

Sobre a base de conhecimento usada no modelo, uma possibilidade seria, ao invés de usar uma base de dados ligados, usar um grande conjunto de mapas conceituais, atualmente disponíveis em vários servidores abertos, para formação de uma grande rede informacional. Essa rede seria consultada tal como ocorre com a base de dados ligados, porém, ao invés de fornecer triplas de RDFs dos dados ligados, ela forneceria proposições dos mapas conceituais. Um problema aparente nessa proposta é de ordem ontológica, pois não haveria a padronização encontrada nos predicados dos dados ligados, mas, frases de ligação das proposições que, normalmente, não seguem um padrão.

Uma área relativamente nova, denominada de *Networked Knowledge Organization Systems/Services* (NKOS), ou Sistemas e Serviços para Organização do Conhecimento em

[79] SKOS: é uma área de trabalho que desenvolve especificações e normas para apoiar o uso de sistemas de organização do conhecimento (*Knowledge Organization System* - KOS), tais como dicionários , esquemas de classificação , sistemas de cabeçalhos de assuntos e taxonomias no âmbito da web semântica . Disponível em : <http://www.w3.org/2004/02/skos/>.

Rede, pode oferecer elementos interessantes para desenvolvimento de projetos para acesso a dados abertos ligados. Segundo o site http://nkos.slis.kent.edu/, NKOS é uma área dedicada à discussão do modelo funcional e de dados em sistemas de classificação, enciclopédias, dicionários, ontologias e serviços de informação interativos em rede que apoiam a recuperação de diversos recursos de informação através da internet. Acredita-se que, munido desse conhecimento e após investigação adequada, seja possível criar recursos diferentes para RI em dados ligados, incrementando os projetos com elementos inovadores.

REFERÊNCIAS

ADAMIC, Lada. *Social network analysis*. E-learning: Coursera, School of Information, University of Michigan, 2013. Disponível em: <https://www.coursera.org/course/sna>. Acesso em: 10 out. 2013.

ALBERT, Réka; BARABÁSI, Albert-László. Statistical mechanics of complex networks. *Reviews of modern physics*, v. 74, n. 1, p. 47, 2002. Disponível em: <http://journals.aps.org/rmp/abstract/10.1103/RevModPhys.74.47>. Acesso em: 20 maio 2016.

ALBERT, Réka; JEONG, Hawoong; BARABÁSI, Albert-László. Internet: Diameter of the World-Wide Web. *Nature*, v. 401, n. 6749, p. 130–131, 9 set. 1999. Disponível em: <http://www.nature.com/nature/journal/v401/n6749/full/401130a0.html>. Acesso em: 24 fev. 2016.

ALMEIDA, Daniela Pereira dos Reis De *et al.* Paradigmas contemporâneos da Ciência da Informação: a recuperação da informação como ponto focal. *Revista Eletrônica Informação e Cognição (Cessada)*, v. 6, n. 1, 2007. Disponível em: <http://www2.marilia.unesp.br/revistas/index.php/reic/article/view/745>. Acesso em: 30 mar. 2016.

ALMEIDA, Gladis Maria de Barcellos. A problemática epistemológica em terminologia: relação entre conceitos. *ALFA: Revista de Linguística*, v. 42, n. 1, 1998. Disponível em: <http://seer.fclar.unesp.br/alfa/article/view/4052>. Acesso em: 7 mar. 2016.

ALMEIDA JÚNIOR, Oswaldo Francisco De. Mediação da informação e múltiplas linguagens. *Tendências da Pesquisa Brasileira em Ciência da Informação*, v. 2, n. 1, 10 ago. 2009. Disponível em: <http://inseer.ibict.br/ancib/index.php/tpbci/article/view/17>. Acesso em: 23 fev. 2016.

ARAÚJO, Carlos Alberto Ávila. O que é Ciência da Informação? *Informação & Informação*, v. 19, n. 1, p. 01–30, 2014. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/15958>. Acesso em: 22 fev. 2016.

ARAÚJO, Eliany Alvarenga De. Equação do impacto informacional: uma proposta paradigmática. In: V ENCONTRO NACIONAL DE PEDQUISA EM CIÊNCIA DA INFORMAÇÃO - ENANCIB, 2003, Belo Horizonte. *Anais...* Belo Horizonte: ECI/UFMG, 2003. Disponível em: <http://enancib.ibict.br/index.php/enancib/venancib/paper/view/2125/1260>. Acesso em: 6 abr. 2016.

ARAÚJO JUNIOR, Rogério Henrique De. *Precisão no processo de busca e recuperação da informação*. Brasília: Thesaurus, 2007.

AUER, Sören *et al.* DBpedia: a nucleus for a web of open data. In: ABERER, KARL *et al.* (Org.). . *The Semantic Web*. Lecture Notes in Computer Science. [S.l.]: Springer Berlin Heidelberg, 2007. p. 722–735. Disponível em: <http://link.springer.com/chapter/10.1007/978-3-540-76298-0_52>. Acesso em: 27 abr. 2016.

AUER, Sören *et al.* Introduction to linked data and its lifecycle on the web. In: RUDOLPH, SEBASTIAN *et al.* (Org.). . *Reasoning web: semantic technologies for intelligent data access*. Lecture Notes in Computer Science. Berlin: Springer Berlin Heidelberg, 2013. p. 1–90. Disponível em: <http://link.springer.com/chapter/10.1007/978-3-642-39784-4_1>. Acesso em: 19 fev. 2016.

AUSUBEL, David Paul. *Educational psychology: a cognitive view*. New York and Toronto: Holt, Rinehart and Winston, 1968.

AUSUBEL, David Paul. *The acquisition and retention of knowledge: a cognitive view*. Dordrecht: Springer Netherlands, 2000. . Acesso em: 16 mar. 2016.

BABIN, Pierre; KOULOUMDJIAN, Marie-France. *Les nouveaux modes de comprendre : la génération de l'audiovisuel et de l'ordinateur*. Paris: Editions du Centurion, 1983.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. *Modern information retrieval: the concepts and technology behind search*. 2. ed. New York: Addison-Wesley, Pearson, 2011.

BAKER, Wayne E.; FAULKNER, Robert R. The Social Organization of Conspiracy: Illegal Networks in the Heavy Electrical Equipment Industry. *American Sociological Review*, v. 58, n. 6, p. 837–860, 1993. Disponível em: <http://www.jstor.org/stable/2095954>. Acesso em: 22 fev. 2016.

BARABÁSI, Albert-László. *Linked: the new science of networks*. Cambridge, Mass: Perseus Pub, 2002.

BARABÁSI, Albert-László. Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, v. 371, n. 1987, p. 20120375–20120375, 18 fev. 2013. Disponível em: <http://rsta.royalsocietypublishing.org/cgi/doi/10.1098/rsta.2012.0375>. Acesso em: 15 fev. 2016.

BARABÁSI, Albert-László. *Network science: graph theory*. Web page: Barabasi site, 2014. Disponível em: <http://barabasi.com/f/625.pdf>. Acesso em: 30 abr. 2016.

BARABÁSI, Albert-László; ALBERT, Réka. Emergence of scaling in random networks. *Science*, PMID: 10521342, v. 286, n. 5439, p. 509–512, 15 out. 1999. Disponível em: <http://science.sciencemag.org/content/286/5439/509>. Acesso em: 24 fev. 2016.

BARAN, Paul. On Distributed Communications Networks. *IEEE Transactions on Communications Systems*, v. 12, n. 1, p. 1–9, mar. 1964. Disponível em: <http://dx.doi.org/10.1109/TCOM.1964.1088883>.

BATISTA, Fábio Ferreira; COSTA, Sely Maria de Souza; ALVARES, Lillian Maria Araújo de Rezende. Gestão do conhecimento : a realização da proposta de Brookes para a Ciência da Informação? In: VIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, out. 2007, Salvador. *Anais...* Salvador: ANCIB, UFBA, out. 2007. Disponível em: <http://repositorio.unb.br/handle/10482/1006>. Acesso em: 30 mar. 2016.

BAUER, Florian; KALTENBÖCK, Martin. *Linked open data: the essentials: a quick start guide for decision makers*. Vienna, Austria: edition mono/monochrom, 2012. Disponível em: <http://www. semantic-web.at/LOD-TheEssentials.pdf>. Acesso em: 12 out. 2015.

BAWDEN, D. Brookes equation: the basis for a qualitative characterization of information behaviours. *Journal of Information Science*, v. 37, n. 1, p. 101–108, 1 fev. 2011. Disponível em: <http://openaccess.city.ac.uk/3130/>. Acesso em: 29 mar. 2016.

BELKIN, Nicholas J. Some(what) grand challenges for information retrieval. *SIGIR Forum*, v. 42, n. 1, p. 47–54, jun. 2008. Disponível em: <http://doi.acm.org/10.1145/1394251.1394261>. Acesso em: 11 abr. 2016.

BELLUZZO, Regina Célia Baptista. O uso de mapas conceituais e mentais como tecnologia de apoio à gestão da informação e da comunicação: uma área interdisciplinar da competência em informação. *Revista Brasileira de Biblioteconomia e Documentação*, São Paulo, v. 2, n. 2, p. 78–89, dez. 2006. Disponível em: <http://rbbd.febab.org.br/rbbd/article/download/19/7>. Acesso em: 4 mar. 2016.

BELLUZZO, Regina Célia Baptista; FERES, Glória Georges; BASSETTO, Clemilton Luis. A competência em informação como um fator crítico de sucesso para a pesquisa na área de ciência da informação: transferência de princípios para reflexão. *Revista EDICIC*, v. 1, n. 1, 14 nov. 2011. Disponível em: <http://www.edicic.org/revista/index.php?journal=RevistaEDICIC&page=article&op=view&path%5B%5D=21>. Acesso em: 22 fev. 2016.

BEPPLER, Fabiano. *Um modelo para recuperação e busca de informação baseado em ontologia e no círculo hermenêutico*. 2008. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento, UFSC, Florianópolis, 2008. Disponível em: <https://repositorio.ufsc.br/handle/123456789/90972>. Acesso em: 8 abr. 2016.

BERNERS-LEE, Tim. *Linked Data*. Web page: W3C, 2006. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 24 fev. 2016.

BERNERS-LEE, Tim. *Long live the web: a call for continued open standards and neutrality*. Web page: Scientific American, 2010. Disponível em: <http://www.scientificamerican.com/article/long-live-the-web/>. Acesso em: 22 fev. 2016.

BERNERS-LEE, Tim *et al.* The semantic web. *Scientific american*, v. 284, n. 5, p. 28–37, 2001. Disponível em: <http://isel2918929391.googlecode.com/svn-history/r347/trunk/RPC/Slides/p01_theSemanticWeb.pdf>. Acesso em: 19 fev. 2016.

BERNERS-LEE, Tim. *Web for real people*. Web page: W3C - Word Wide Web Consortium, 2005. Disponível em: <https://www.w3.org/2005/Talks/0511-keynote-tbl/>. Acesso em: 30 abr. 2016.

BERNERS-LEE, Tim; O'HARA, Kieron. The read–write Linked Data Web. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, PMID: 23419858, v. 371, n. 1987, p. 20120513, 28 mar. 2013. Disponível em: <http://rsta.royalsocietypublishing.org/content/371/1987/20120513>. Acesso em: 24 fev. 2016.

BICALHO, Lucinéia Maria. As relações interdisciplinares refletidas na Ciência da Informação. *Perspectivas em Ciência da Informação*, v. 15, n. 1, p. 309–309, abr. 2010. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362010000100018&lang=pt>. Acesso em: 1 out. 2012.

BILETZKI, Anat; MATAR, Anat. Ludwig Wittgenstein. In: ZALTA, EDWARD N. (Org.). . *The Stanford Encyclopedia of Philosophy*. Spring 2014 ed. [S.l: s.n.], 2014. . Disponível em: <http://plato.stanford.edu/archives/spr2014/entries/wittgenstein/>. Acesso em: 2 mar. 2016.

BISPO, Carlos Alberto Ferreira; CASTANHEIRA, Luiz Batista; SOUZA FILHO, Oswaldo Melo. *Introdução à Lógica Matemática*. São Paulo: Cengage Learning, 2011.

BIZER, Christian; HEATH, Tom; BERNERS-LEE, Tim. Linked data - the story so far: *International Journal on Semantic Web and Information Systems*, v. 5, n. 3, p. 1–22, 33 2009. Disponível em: <http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/jswis.2009081901>. Acesso em: 19 fev. 2016.

BRAGA, Katia Soares. Aspectos relevantes para a seleção de metodologia adequada à pesquisa social em Ciência da Informação. In: MUELLER, SUZANA P. M. *Métodos para a pesquisa em Ciência da Informação*. Brasília: Tesaurus, 2007. p. 17–38.

BRASIL. Constituição (1988). Constituição da República Federativa do Brasil. Brasília, DF, Senado, 1998.

BRASIL. Decreto nº 7.724, de 16 de maio de 2012. Regulamenta a Lei no 12.527, de 18 de novembro de 2011, que dispõe sobre o acesso a informações previsto no inciso XXXIII do caput do art. 5o, no inciso II do § 3o do art. 37 e no § 2o do art. 216 da Constituição. Diário oficial [da] República Federativa do Brasil. Brasília, maio 2012.

BRASIL. Lei nº 12.527, de 18 de novembro de 2011. Regula o acesso a informações... Diário oficial [da] República Federativa do Brasil. Brasília, 18 nov. 2011. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm>.

BROOKES, Bertram C. The foundations of information science - part III - quantitative aspects: objective maps and subjective landscapes. *Journal of Information Science*, v. 2, n. 6, p. 269–275, 1 dez. 1980a. Disponível em: <http://jis.sagepub.com/content/2/6/269>. Acesso em: 7 abr. 2016.

BROOKES, Bertram C. The foundations of information science - part II - quantitative aspects: classes of things and the challenge of human individuality. *Journal of Information Science*, v. 2, n. 5, p. 209–221, 1 out. 1980b. Disponível em: <http://jis.sagepub.com/content/2/5/209>. Acesso em: 7 abr. 2016.

BROOKES, Bertram C. The foundations of information science - part I - philosophical aspects. *Journal of Information Science*, London, v. 2, n. 3-4, p. 125–133, 1 jun. 1980c. Disponível em: <http://jis.sagepub.com/content/2/3-4/125>. Acesso em: 29 mar. 2016.

BROOKES, Bertram C. The foundations of information science - part IV - information science: the changing paradigm. *Journal of Information Science*, v. 3, n. 1, p. 3–12, 1 fev. 1981. Disponível em: <http://jis.sagepub.com/content/3/1/3>. Acesso em: 7 abr. 2016.

BUCHANAN, Mark. *Nexus: small worlds and the groundbreaking science of networks*. New York: W W Norton & Company, 2002.

BUCKLAND, Michael; GEY, Fredric. The relationship between Recall and Precision. *Journal of the American Society for Information Science*, v. 45, n. 1, p. 12–19, 1 jan. 1994.

Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L/abstract>. Acesso em: 7 jul. 2016.

BUCKLAND, Michael K. Information as thing. *Journal of the American Society for Information Science*, v. 42, n. 5, p. 351, jun. 1991. Disponível em: <http://inls151f14.web.unc.edu/files/2014/08/buckland1991-informationasthing.pdf>. Acesso em: 1 abr. 2016.

BUCKLAND, Michael K. *et al.* OASIS: A front-end for prototyping catalog enhancements. *Library Hi Tech*, v. 10, n. 4, p. 7–22, 1 abr. 1992. Disponível em: <http://www.emeraldinsight.com/doi/abs/10.1108/eb047860>. Acesso em: 9 jul. 2016.

BURKHARD, Remo Aslak. Towards a framework and a model for knowledge visualization: synergies between information and knowledge visualization. In: TERGAN, SIGMAR-OLAF; KELLER, TANJA (Org.). . *Knowledge and Information Visualization*. Lecture Notes in Computer Science. [S.l.]: Springer Berlin Heidelberg, 2005. p. 238–255. Disponível em: <http://link.springer.com/chapter/10.1007/11510154_13>. Acesso em: 4 jul. 2016.

BUSH, Vannevar. As We May Think. *The Atlantic*, jul. 1945. Disponível em: <http://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/>. Acesso em: 29 fev. 2016.

CAMPOS, Maria Luiza de Almeida; SOUZA, Rosali Fernandez De; CAMPOS, Maria Luiza Machado. Organização de unidades de conhecimento em hiperdocumentos: o modelo conceitual como espaço comunicacional para a realização da autoria. *Ciência da Informação*, v. 32, n. 2, 22 ago. 2003. Disponível em: <http://revista.ibict.br/index.php/ciinf/article/view/111>.

CAÑAS, Alberto J. *et al.* Concept Maps: Integrating Knowledge and Information Visualization. In: TERGAN, SIGMAR-OLAF; KELLER, TANJA (Org.). . *Knowledge and Information Visualization*. Lecture Notes in Computer Science. Berlin: Springer Berlin Heidelberg, 2005. p. 205–219. Disponível em: <http://cmap.ihmc.us/publications/ResearchPapers/ConceptMapsIntegratingKnowInfVisual.pdf>. Acesso em: 30 jun. 2016.

CAÑAS, Alberto J. *et al.* Mining the web to suggest concepts during concept map construction. In: FIRST INT. CONFERENCE ON CONCEPT MAPPING, set. 2004, Pamplona, Spain. *Anais*... Pamplona, Spain: Dirección de Publicaciones de la Universidad Pública de Navarra, set. 2004. Disponível em: <http://eprint.ihmc.us/91/1/cmc2004-284.pdf>. Acesso em: 4 mar. 2016.

CAÑAS, Alberto J.; NOVAK, Joseph Donald; REISKA, Priit. How good is my concept map? am I a good cmapper? *Knowledge Management & E-Learning*, v. 7, n. 1, p. 6–19, mar. 2015. Disponível em: <http://www.kmel-journal.org/ojs/index.php/online-publication/article/viewFile/407/244>. Acesso em: 17 mar. 2016.

CANFORA, Gerardo; CERULO, Luigi. A taxonomy of information retrieval models and tools. *Journal of Computing and Information Technology*, v. 12, n. 3, p. 175, 2004. Disponível em: <http://cit.srce.unizg.hr/index.php/CIT/article/view/1546>. Acesso em: 13 abr. 2016.

CAPURRO, Rafael. Epistemologia e ciência da informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 2003, Belo Horizonte. *Anais*... Belo Horizonte: Associação Nacional de Pesquisa e Pós-Graduação em Ciência da Informação, 2003. Disponível em: <http://www.capurro.de/enancib_p.htm>. Acesso em: 23 fev. 2016.

CAPURRO, Rafael; HJØRLAND, Birger. The concept of information. *Annual Review of Information Science and Technology*, v. 37, p. 343–411, 2003. Disponível em: <http://fiz1.fh-potsdam.de/volltext/stuttgart/04058.html>. Acesso em: 23 fev. 2016.

CASADO, Elías Sanz. *Manual de estudios de usuarios*. Madrid; Madrid: Fundación Germán Sánchez Ruipérez ; Pirámide, 1994.

CASE, Donald O. Information behavior. *Annual Review of Information Science and Technology*, v. 40, n. 1, p. 293–327, 1 jan. 2006. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/aris.1440400114/abstract>. Acesso em: 24 fev. 2016.

CASTELLS, Manuel; CASTELLS, Manuel. *The rise of the network society*. 2nd ed ed. Oxford ; Malden, Mass: Blackwell Publishers, 2000. (Information age, v. 1).

CHAMPCLAUX, Yaël; DKAKI, Taoufiq; MOTHE, Josiane. An information retrieval models taxonomy based on an analogy between cognitive science and information retrieval. 2010, Toulouse, France. *Anais*... Toulouse, France: [s.n.], 2010. Disponível em: <https://www.irit.fr/publis/SIG/2010_VSST_CDM.pdf>. Acesso em: 13 abr. 2016.

CHAYES, Jennifer. Mathematics of Web science: structure, dynamics and incentives. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, PMID: 23419846, v. 371, n. 1987, p. 20120377, 28 mar. 2013. Disponível em: <http://rsta.royalsocietypublishing.org/content/371/1987/20120377>. Acesso em: 25 fev. 2016.

CHEN, Chaomei. *Mapping scientific frontiers: the quest for knowledge visualization*. 2. ed. London: Springer Science & Business Media, 2013.

CLEVERDON, Cyril W. On the inverse relationship of recall and precision. *Journal of Documentation*, v. 28, n. 3, p. 195–201, 1 mar. 1972. Disponível em: <http://www.emeraldinsight.com/doi/abs/10.1108/eb026538>. Acesso em: 8 jul. 2016.

CLEVERDON, Cyril W. *Report on the testing and analysis of an investigation into the comparative efficiency of indexing systems*. Cranfield: Aslib, 1962.

CONCEIÇÃO, Simone C. O.; SAMUEL, Anita; BINIECKI, Susan M. Yelich. Application of concept maps for conducting research. In: SIXTH INTERNATIONAL CONFERENCE ON CONCEPT MAPPING, 2014, Santos-SP. *Anais*... Santos-SP: USP, IHMC, 2014. p. 63–70. Disponível em: <http://cmc.ihmc.us/cmc2014Program.html>. Acesso em: 3 abr. 2016.

CRESWELL, John W. *Research design: qualitative, quantitative, and mixed methods approaches*. [S.l.]: SAGE Publications, 2009.

CRISTOVÃO, Henrique Monteiro *et al.* O ensino de mapas conceituais a alunos-professores em um curso de pós-graduação lato sensu ofertado a distância. In: SIXTH INTERNATIONAL CONFERENCE ON CONCEPT MAPPING, 2014, Santos-SP. *Anais*...

Santos-SP: USP, IHMC, 2014. p. 730–733. Disponível em: <http://cmc.ihmc.us/cmc2014Program.html>. Acesso em: 3 abr. 2016.

CRISTOVÃO, Henrique Monteiro. Uma experiência com o editor de textos: hipertexto e revisão. In: XI SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2000, Maceió. *Anais*... Maceió: UFAL-SBC, 2000. p. 437–439.

CURY, Davidson; PERIN, Wagner Andrade; SANTOS JUNIOR, Isaura Alcina Martins. CMPAAS - a platform of services for construction and handling of concept maps. In: SIXTH INTERNATIONAL CONFERENCE ON CONCEPT MAPPING, 2014, Santos-SP. *Anais*... Santos-SP: USP, IHMC, 2014. p. 107–115. Disponível em: <http://cmc.ihmc.us/cmc2014Program.html>. Acesso em: 3 abr. 2016.

CYGANIAK, Richard *et al. RDF 1.1 concepts and abstract syntax*. Web page: W3C, 2014. Disponível em: <http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225>. Acesso em: 24 fev. 2016.

DAHLBERG, Ingetraut. Teoria do conceito. *Ciência da informação*, v. 7, n. 2, 1978. Disponível em: <http://revista.ibict.br/index.php/ciinf/article/viewArticle/1680>. Acesso em: 19 fev. 2016.

DING, Li; PERISTERAS, Vassilios; HAUSENBLAS, Michael. Linked open government data. *IEEE Intelligent Systems*, v. 27, n. 3, p. 11–15, 2012. Disponível em: <https://www.computer.org/csdl/mags/ex/2012/03/mex2012030011-abs.html>. Acesso em: 31 mar. 2016.

DUQUE, Cláudio Gottschalg. *SIRILICO - uma proposta para um sistema de recuperação de informação baseado em teorias da lingüística computacional e ontologia*. 2005. Tese (Doutorado em Ciência da Informação) - Programa de Pós-Graduação em Ciência da Informação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

DUTRA, Ítalo Modesto; FAGUNDES, Léa da Cruz; CAÑAS, Alberto J. Uma proposta de uso dos mapas conceituais para um paradigma construtivista da formação de professores a distância. In: X WIE - WORKSHOP SOBRE INFORMÁTICA NA ESCOLA, 2004, Salvador. *Anais*... Salvador: [s.n.], 2004. Disponível em: <http://www.pead.faced.ufrgs.br/sites/tutoriais/trilha-antiga/mapas_conceituais/documentos/mapas_italo_lea_canas.pdf>. Acesso em: 4 mar. 2016.

EPPLER, Martin J; BURKHARD, Remo Aslak. *Knowledge visualization: towards a new discipline and its field of application*. Lugano, Switzerland: Università della Svizzera italiana, 2004. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.134.6040>. Acesso em: 30 maio 2016.

ERDŐS, Paul; RÉNYI, A. On random graphs I. *Publicationes Mathematicae (Debrecen)*, v. 6, p. 290–297, 1959. Disponível em: <http://ftp.math-inst.hu/~p_erdos/1959-11.pdf>.

FERNEDA, Edberto. *Recuperação da informação: análise sobre a contribuição da Ciência da Computação para a Ciência da Informação*. 2003. Tese (Doutorado em Ciências da Comunicação) - Escola de Comunicação e Artes da Universidade Federal de São Paulo, São Paulo, 2003. Disponível em: <http://www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/publico/Tese.pdf>. Acesso em: 31 mar. 2016.

FISHER, Karen E.; JULIEN, Heidi. Information behavior. *Annual Review of Information Science and Technology*, v. 43, n. 1, p. 1–73, 1 jan. 2009. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/aris.2009.1440430114/abstract>. Acesso em: 24 fev. 2016.

FLORIDI, Luciano. *On defining library and information science as applied philosophy of information*. [S.l: s.n.], 2002.

FONSECA, Maria Odila Kahl. *Arquivologia e ciência da informação*. 1. ed ed. Rio de Janeiro: FGV, 2005.

FRANCO, Augusto De. *A rede*. São Paulo: Escola de Redes, 2012. Disponível em: <http://pt.slideshare.net/augustodefranco/fluzz-srie-completa>. Acesso em: 23 fev. 2016.

GAINES, Brian R.; SHAW, Mildred L. G. Concept maps as hypermedia components. *International Journal of Human-Computer Studies*, Duluth-MN, USA, v. 43, n. 3, p. 323–361, set. 1995. Disponível em: <http://dx.doi.org/10.1006/ijhc.1995.1049>. Acesso em: 10 mar. 2016.

GASQUE, Kelley Cristine Gonçalves Dias; COSTA, Sely Maria de Souza. Evolução teórico-metodológica dos estudos de comportamento informacional de usuários. *Ciência da Informação*, v. 39, n. 1, p. 21–32, 2010. Disponível em: <http://www.scielo.br/pdf/ci/v39n1/v39n1a02>. Acesso em: 23 fev. 2016.

GIL, Antonio Carlos. *Como elaborar projetos de pesquisa*. São Paulo: Atlas, 2002.

GLEICK, James. *The information: a history, a theory, a flood*. [S.l.]: Knopf Doubleday Publishing Group, 2011.

GOH, Kwang-Il *et al.* The human disease network. *Proceedings of the National Academy of Sciences*, PMID: 17502601, v. 104, n. 21, p. 8685–8690, 22 maio 2007. Disponível em: <http://www.pnas.org/content/104/21/8685>. Acesso em: 22 fev. 2016.

GORDON, Michael; KOCHEN, Manfred. Recall-precision trade-off: a derivation. *Journal of the American Society for Information Science*, v. 40, n. 3, p. 145–151, 1 maio 1989. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(198905)40:3<145::AID-ASI1>3.0.CO;2-I/abstract>. Acesso em: 8 jul. 2016.

GRANOVETTER, Mark. The strength of weak ties: a network theory revisited. *Sociological Theory*, v. 1, n. 1983, p. 201–233, 1983. Disponível em: <https://sociology.stanford.edu/sites/default/files/publications/the_strength_of_weak_ties_and_exch_w-gans.pdf>. Acesso em: 6 maio 2016.

GRANOVETTER, Mark S. The strength of weak ties. *American Journal of Sociology*, v. 78, n. 6, p. 1360–1380, 1 maio 1973. Disponível em: <http://www.journals.uchicago.edu/doi/10.1086/225469>. Acesso em: 6 maio 2016.

GRAUDINA, Vita; GRUNDSPENKIS, Janis. Concept map generation from OWL ontologies. In: THIRD INTERNATIONAL CONFERENCE ON CONCEPT MAPPING, 2008, Tallinn, Estonia and Helsinki, Finland. *Anais*... Tallinn, Estonia and Helsinki, Finland: Tallinn University, IHMC, University of Helsinki, 2008. p. 263–270. Disponível em: <http://cmc.ihmc.us/cmc2008papers/cmc2008-p263.pdf>. Acesso em: 27 abr. 2016.

GREEN, Sheryl. Applying concept maps to analyse the level of sustainable development awareness held by policy makers in Malta. In: FIFTH INTERNATIONAL CONFERENCE ON CONCEPT MAPPING, 17 set. 2012, Valleta, Malta. *Anais*... Valleta, Malta: University of Malta, IHMC, 17 set. 2012. Disponível em: <http://cmc.ihmc.us/cmc2012/CMC2012Program.html>. Acesso em: 22 fev. 2016.

GUÉRET, Christophe *et al.* Assessing linked data mappings using network measures. ESWC'12, 2012, Berlin, Heidelberg. *Anais*... Berlin, Heidelberg: Springer-Verlag, 2012. p. 87–102. Disponível em: <http://dx.doi.org/10.1007/978-3-642-30284-8_13>. Acesso em: 19 fev. 2016.

GURRIN, Cathal *et al.* Recent developments in information retrieval. In: HUTCHISON, DAVID *et al.* (Org.). . *Advances in Information Retrieval*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. v. 5993. p. 1–9. Disponível em: <http://link.springer.com/10.1007/978-3-642-12275-0>. Acesso em: 11 abr. 2016.

HAGEDORN, Kat. *The information architecture glossary*. Web page: Argus Center for Information Arquitecture, 2000. Disponível em: <http://argus-acia.com/white_papers/iaglossary.html>. Acesso em: 31 mar. 2016.

HAUSENBLAS, Michael; KIM, James G. *Five star open data*. Web page: LATC, 2015. Disponível em: <http://5stardata.info/en/>. Acesso em: 24 fev. 2016.

HEATH, Tom. *Linked data? Web of data? Semantic web? WTF?* Web page: Tom Heath's Displacement Activities, 2009. Disponível em: <http://tomheath.com/blog/2009/03/linked-data-web-of-data-semantic-web-wtf/>. Acesso em: 1 maio 2016.

HEATH, Tom; BIZER, Christian. Linked data: evolving the web into a global data space. *Synthesis Lectures on the Semantic Web: Theory and Technology*, v. 1, n. 1, p. 1–136, 9 fev. 2011. Disponível em: <http://www.morganclaypool.com/doi/abs/10.2200/S00334ED1V01Y201102WBE001>. Acesso em: 19 fev. 2016.

HEDDEN, Heather. Controlled vocabularies, thesauri, and taxonomies. *The Indexer*, v. 26, n. 1, p. 33–34, 1 mar. 2008. Disponível em: <http://www.theindexer.org/online.htm>. Acesso em: 19 jun. 2016.

HEIM, Philipp *et al.* RelFinder: revealing relationships in RDF knowledge bases. In: CHUA, TAT-SENG *et al.* (Org.). . *Semantic Multimedia*. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. p. 182–187. Disponível em: <https://www.uni-due.de/~s400268/RelFinder-SAMT09.pdf>. Acesso em: 4 jul. 2016.

HEIM, Philipp; ERTL, Thomas; ZIEGLER, Jürgen. Facet Graphs: complex semantic querying made easy. In: AROYO, LORA *et al.* (Org.). . *The semantic web: research and applications*. Lecture Notes in Computer Science. Berlin: Springer Berlin Heidelberg, 2010. p. 288–302. Disponível em: <http://www.sfb716.uni-stuttgart.de/uploads/tx_vispublications/eswc10-heimErtlZiegler.pdf>. Acesso em: 4 maio 2016.

HENDLER, James *et al.* Web science: an interdisciplinary approach to understanding the web. *Communications of the ACM*, v. 51, n. 7, p. 60, 1 jul. 2008. Disponível em: <http://portal.acm.org/citation.cfm?doid=1364782.1364798>. Acesso em: 17 mar. 2016.

HIDALGO, C. A. *et al.* The Product Space Conditions the Development of Nations. *Science*, PMID: 17656717, v. 317, n. 5837, p. 482–487, 27 jul. 2007. Disponível em: <http://science.sciencemag.org/content/317/5837/482>. Acesso em: 24 fev. 2016.

HJØRLAND, Birger. The foundation of the concept of relevance. *Journal of the American Society for Information Science and Technology*, v. 61, n. 2, p. 217–237, 1 fev. 2010. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/asi.21261/abstract>. Acesso em: 16 fev. 2016.

HOFFMAN, Robert; BEACH, Jameson. Lessons learned across a decade of knowledge modeling. *JETT*, v. 4, n. 1, p. 85–95, 2013. Disponível em: <http://dialnet.unirioja.es/servlet/articulo?codigo=4264824>. Acesso em: 22 fev. 2016.

HOOK, Peter A.; BÖRNER, Katy. Educational knowledge domain visualizations: tools to navigate, understand, and internalize the structure of scholarly knowledge and expertise. In: SPINK, AMANDA; COLE, CHARLES (Org.). . *New directions in cognitive information retrieval*. The Information Retrieval Series. Amsterdam: Springer Netherlands, 2005. p. 187–208. Disponível em: <http://cns.iu.edu/images/pub/2005-hook-educknow.pdf>. Acesso em: 29 jun. 2016.

HOWE, Jeff. The Rise of Crowdsourcing. *Wired Magazine*, v. 14, n. 6, 1 jun. 2006. Disponível em: <http://www.wired.com/2006/06/crowds/>. Acesso em: 24 fev. 2016.

INGWERSEN, Peter; JÄRVELIN, Kalervo. *The turn: integration of information seeking and retrieval in context*. Berlin, Heidelberg: Springer-Verlag, 2005. v. 18. Disponível em: <http://link.springer.com/10.1007/1-4020-3851-8>. Acesso em: 7 jul. 2016. (The Information Retrieval Series).

KADUSHIN, Charles. Introduction to social network theory. *Boston, MA*, 2004. Disponível em: <http://melander335.wdfiles.com/local--files/reading-history/kadushin.pdf>. Acesso em: 24 fev. 2016.

KELLER, Tanja; TERGAN, Sigmar-Olaf. Visualizing knowledge and information: an introduction. In: KELLER, TANJA; TERGAN, SIGMAR-OLAF. *Knowledge and information visualization*. Berlin: Springer Berlin Heidelberg, 2005. p. 1–23. Disponível em: <http://ldt.stanford.edu/~educ39105/paul/articles_2005/visualizing%20knowledge%20and%20information.pdf>. Acesso em: 29 jun. 2016.

KELLY, Diane. Methods for evaluating interactive information retrieval systems with users. *Found. Trends Inf. Retr.*, v. 3, n. 1—2, p. 1–224, jan. 2009. Disponível em: <http://dl.acm.org/citation.cfm?id=1618302>. Acesso em: 10 abr. 2016.

KELLY, Diane; SUGIMOTO, Cassidy R. A systematic review of interactive information retrieval evaluation studies, 1967–2006. *Journal of the American Society for Information Science and Technology*, v. 64, n. 4, p. 745–770, 1 abr. 2013. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/asi.22799/abstract>. Acesso em: 10 abr. 2016.

KÉPÉKLIAN, Gabriel; CURÉ, Olivier; BIHANIC, Laurent. From the web of documents to the linked data. In: ZIMÁNYI, ESTEBAN; KUTSCHE, RALF-DETLEF (Org.). . *Business Intelligence*. Lecture Notes in Business Information Processing. Cham, Switzerland: Springer International Publishing, 2015. p. 60–87. Disponível em: <http://link.springer.com/chapter/10.1007/978-3-319-17551-5_3>. Acesso em: 27 abr. 2016.

KLEINBERG, Jon. Analysis of large-scale social and information networks. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, PMID: 23419847, v. 371, n. 1987, p. 20120378, 28 mar. 2013. Disponível em: <http://rsta.royalsocietypublishing.org/content/371/1987/20120378>. Acesso em: 6 maio 2016.

KOWATA, Juliana H.; CURY, Davidson; BOERES, MC Silva. A review of semi-automatic approaches to build concept maps. 2010, Viña del Mar, Chile. *Anais*... Viña del Mar, Chile: Lom Ediciones S.A., 2010. Disponível em: <http://cmc.ihmc.us/cmc/CMCProceedings.html>. Acesso em: 25 fev. 2016.

KUROPKA, Dominik. *Modelle zur repräsentation natürlichsprachlicher dokumente: ontologie-basiertes information-filtering und -retrieval mit relationalen datenbanken*. Berlin: Logos-Verl, 2004. (Advances in information systems and management science, 10).

LANCASTER, F. W. *Indexing and abstracting in theory and practice*. 3. ed. Champaign, IL, USA: University of Illinois, 2003.

LANZING, Jan. What is concept mapping? *The concept mapping homepage*. Web page: [s.n.], 1997. . Disponível em: <http://users.edte.utwente.nl/lanzing/cm_home.htm>. Acesso em: 10 mar. 2016.

LE COADIC, Yves-françois. *A Ciência da informação*. Brasilia: Briquet de Lemos Livros, 1996.

LEHMANN, Jens *et al.* DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web Journal*, v. 6, n. 2, p. 167–195, 2015. Disponível em: <http://jens-lehmann.org/files/2014/swj_dbpedia.pdf>.

LEHMANN, Jens; SCHÜPPEL, Jörg; AUER, Sören. Discovering unknown connections-the DBpedia relationship finder. *CSSW*, v. 113, p. 99–110, 2007. Disponível em: <http://www.informatik.uni-leipzig.de/~auer/publication/relfinder.pdf>. Acesso em: 4 jul. 2016.

LEVY, Pierre. *As tecnologias da inteligência*. Tradução Carlos Irineu Da Costa. [S.l.]: Editora 34, 1993.

LÉVY, Pierre. *O que é o virtual?* Tradução Carlos Irineu Da Costa. São Paulo: Editora 34, 1996.

LÉVY, Pierre; AUTHIER, Michel. *As árvores de conhecimentos*. Tradução Monica M. Seineman. São Paulo: Escuta, 1995.

LEVY, Pierre; COSTA, Carlos Irineu Da. *Cibercultura*. [S.l.]: Editora 34, 1999.

LIEW, Anthony. DIKIW: data, information, knowledge, intelligence, wisdom and their interrelationships. *Business Management Dynamics*, v. 2, n. 10, p. 49–62, abr. 2013. Disponível em: <http://bmdynamics.com/recent_issue.php?id=25>. Acesso em: 31 mar. 2016.

LIEW, Anthony. Understanding data, information, knowledge and their inter-relationships. *Journal of Knowledge Management Practice*, v. 8, n. 2, jun. 2007. Disponível em: <https://www.tlainc.com/articl134.htm>. Acesso em: 31 mar. 2016.

LIMA, Gercina Ângela Borém de Oliveira. Mapa conceitual como ferramenta para organização do conhecimento em sistema de hipertextos e seus aspectos cognitivos. *Perspectivas em Ciência da Informação*, v. 9, n. 2, 2004. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/355>. Acesso em: 9 mar. 2016.

LIMA, Gercina Ângela Borém de Oliveira. Modelo hipertextual-MHTX: um modelo para organização hipertextual de documentos. In: VI ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 2005, Florianópolis. *Anais*... Florianópolis: IBICT, 2005. Disponível em: <http://enancib.ibict.br/index.php/enancib/vienancib/schedConf/presentations>. Acesso em: 9 mar. 2016.

LIU, Yi. *Graph-based learning models for information retrieval: a survey*. [S.l: s.n.], 2006. Disponível em: <http://www.cse.msu.edu/~rongjin/semisupervised/graph.pdf>. Acesso em: 14 abr. 2016.

LOGAN, Robert. *What is information?: propagating organization in the biosphere, symbolosphere, technosphere and econosphere*. [S.l.]: DEMO Publishing, 2014.

LOHMANN, Steffen *et al.* The RelFinder user interface: interactive exploration of relationships between objects of interest. IUI '10, 2010, New York, NY, USA. *Anais*... New York, NY, USA: ACM, 2010. p. 421–422. Disponível em: <http://doi.acm.org/10.1145/1719970.1720052>. Acesso em: 4 maio 2016.

MANUAL DOS DADOS ABERTOS: desenvolvedores. São Paulo: Comitê Gestor da Internet no Brasil, 2011. . Disponível em: <http://www.w3c.br/pub/Materiais/PublicacoesW3C/manual_dados_abertos_desenvolvedores_web.pdf>.

MARTELETO, Regina Maria. Análise de redes sociais – aplicação nos estudos de transferência da informação. *Ciência da Informação*, v. 30, n. 1, 12 jun. 2001. Disponível em: <http://revista.ibict.br/ciinf/index.php/ciinf/article/view/226>.

MARTIN, P.; EKLUND, P. W. Knowledge retrieval and the World Wide Web. *IEEE Intelligent Systems and their Applications*, v. 15, n. 3, p. 18–25, maio 2000. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.128.2803&rep=rep1&type=pdf>. Acesso em: 19 abr. 2016.

MATHEUS, Renato Fabiano; SILVA, Antonio Braz de Oliveira E. Análise de redes sociais como método para a Ciência da Informação. *DataGramaZero - Revista de Ciência da Informação*, v. 7, n. 2, 1 abr. 2006. Disponível em: <http://www.dgz.org.br/abr06/Art_03.htm>. Acesso em: 24 fev. 2016.

MCLINDEN, Daniel. Concept maps as network data: analysis of a concept map using the methods of social network analysis. *Evaluation and Program Planning*, v. 36, n. 1, p. 40–48, fev. 2013. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0149718912000456>. Acesso em: 25 fev. 2016.

MEADOW, Charles T. *et al. Text Information Retrieval Systems*. 3. ed. Amsterdam: Academic Press, 2007.

MENEZES, Paulo Blauth. *Matemática discreta para computação e informática*. Porto Alegre: Bookman, 2013.

MEYER, Robert. Knowledge visualization. *Trends in Information Visualization*, v. 23, 2010. Disponível em: <http://www.mmi.ifi.lmu.de/pubdb/publications/pub/baur2010infovisHS/baur2010infovisHS.pdf#page=31>. Acesso em: 30 jun. 2016.

MIKA, Peter. *Social networks and the semantic web*. Boston, MA: Springer US, 2007. v. 5. Disponível em: <http://link.springer.com/10.1007/978-0-387-71001-3>. Acesso em: 22 fev. 2016. (Semantic Web and Beyond).

MILGRAM, Stanley. The small world problem. *Psychology Today*, v. 1, n. 1, p. 60–67, maio 1967. Disponível em: <http://measure.igpp.ucla.edu/GK12-SEE-LA/Lesson_Files_09/Tina_Wey/TW_social_networks_Milgram_1967_small_world_problem.pdf>.

MINAYO, Maria Cecilia de Souza. *Pesquisa social: teoria, método e criatividade*. 18. ed. Petrópolis: Vozes, 2001.

MORAES, Marielle Barros De. A Ciência da Informação nos caminhos do contemporâneo. *PontodeAcesso*, v. 7, n. 2, p. 2–24, 11 set. 2013. Disponível em: <http://www.portalseer.ufba.br/index.php/revistaici/article/view/5199>. Acesso em: 30 mar. 2016.

MOREIRA, Marco Antonio. *Mapas conceituais e aprendizagem significativa, organizadores prévios, mapas conceituais, diagramas V e unidades de ensino potencialmente significantes*. Web page: PUCPR, 2013. Disponível em: <http://paginas.uepa.br/erasnorte2013/images/sampledata/figuras/aprend_%20signif_%20org_prev_mapas_conc_diagr_v_e_ueps.pdf#page=41>. Acesso em: 4 mar. 2016.

MOREIRA, Marco Antonio; MASINI, Elcie F. Salzano. *Aprendizgem significativa: a teoria da David Ausubel*. São Paulo: Moraes, 1982.

MOTTA, Dilza Fonseca Da. *Método relacional como nova abordagem para a construção de tesauros*. São Paulo: SENAC, 1987.

NARANJO, Luis; KAUFFMANN, Erick; FERRÁNDEZ, Antonio. A Taxonomy for information retrieval models based on the uncertainty level. *NOOS*, v. 4, fev. 2014. Disponível em: <http://www.revistanoos.co/volumen-4-2/>. Acesso em: 13 abr. 2016.

NASCIMENTO, Denise Morado. A abordagem sócio-cultural da informação. *Informação & Sociedade: Estudos*, v. 16, n. 2, 2006. Disponível em: <http://www.ies.ufpb.br/ojs/index.php/ies/article/view/477>. Acesso em: 30 mar. 2016.

NATIONAL RESEARCH COUNCIL. *Network Science*. [S.l: s.n.], 2005. Disponível em: <https://www.nap.edu/catalog/11516/network-science>. Acesso em: 20 set. 2016.

NAZARIO, Debora Cabral; SILVA, Paulo Fernando Da; ROVER, Aires José. Avaliação da qualidade da informação disponibilizada no Portal da Transparência do Governo Federal. *Revista Democracia Digital e Governo Eletrônico*, v. 1, n. 6, 15 jun. 2012. Disponível em:

<http://www.buscalegis.ccj.ufsc.br/revistas/index.php/observatoriodoegov/article/view/34154>. Acesso em: 24 fev. 2016.

NEILL, S. D. Brookes, Popper, and objective knowledge. *Journal of Information Science*, v. 4, n. 1, p. 33–39, 1 jan. 1982. Disponível em: <http://jis.sagepub.com/content/4/1/33>. Acesso em: 7 abr. 2016.

NEWMAN, M. E. J. *Networks: an introduction*. Oxford ; New York: Oxford University Press, 2010.

NEWMAN, M. E. J. The structure and function of complex networks. *SIAM Review*, arXiv: cond-mat/0303516, v. 45, n. 2, p. 167–256, jan. 2003. Disponível em: <http://arxiv.org/abs/cond-mat/0303516>. Acesso em: 19 abr. 2016.

NOOY, Wouter De; MRVAR, Andrej; BATAGELJ, Vladimir. *Exploratory social network analysis with Pajek*. Rev. and expanded 2nd ed ed. England ; New York: Cambridge University Press, 2011. (Structural analysis in the social sciences, 34).

NOVAK, Joseph Donald. *A theory of education*. Ithaca, N.Y.: Cornell University Press, 1977. Disponível em: <http://catalog.hathitrust.org/Record/000252496>. Acesso em: 21 fev. 2016.

NOVAK, Joseph Donald. A theory of education: meaningful learning underlies the constructive integration of thinking, feeling, and acting leading to empowerment for commitment and responsability. *Aprendizagem Significativa em Revista/Meaningful Learning Review*, Porto Alegre, v. 1, n. 2, p. 1–14, 2011. Disponível em: <http://www.if.ufrgs.br/asr/?go=artigos&idEdicao=2#>. Acesso em: 4 mar. 2016.

NOVAK, Joseph Donald. *Learning, creating, and using knowledge: concept maps as facilitative tools in schools and corporations*. [S.l.]: Taylor & Francis, 2010.

NOVAK, Joseph Donald; CAÑAS, A. J. The universality and ubiquitousness of concept maps. In: FOURTH INTERNATIONAL CONFERENCE ON CONCEPT MAPPING, 2010, Viña del Mar, Chile. *Anais*... Viña del Mar, Chile: Lom Ediciones S.A., 2010. Disponível em: <http://cmc.ihmc.us/cmc/CMCProceedings.html>. Acesso em: 21 fev. 2016.

NOVAK, Joseph Donald; CAÑAS, Alberto J. Theoretical origins of concept maps, how to construct them, and uses in education. *Reflecting Education*, v. 3, n. 1, p. 29–42, 27 nov. 2007. Disponível em: <http://www.reflectingeducation.net/index.php/reflecting/article/view/41>. Acesso em: 4 mar. 2016.

NOVAK, Joseph Donald; CAÑAS, Alberto J. The theory underlying concept maps and how to construct and use them. 2008. Disponível em: <http://cmap.ihmc.us/Publications/ResearchPapers/TheoryUnderlyingConceptMaps.pdf>. Acesso em: 22 fev. 2016.

NOVAK, Joseph Donald; GOWIN, D. Bob. *Learning how to learn*. Cambridge, New York: Cambridge University Press, 1984.

ODONNELL, Angela. Searching for information in knowledge maps and texts. *Contemporary Educational Psychology*, v. 18, n. 2, p. 222–239, Abril 1993. Disponível em:

<http://www.sciencedirect.com/science/article/pii/S0361476X83710180>. Acesso em: 22 fev. 2016.

ONNELA, J.-P. *et al.* Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, PMID: 17456605, v. 104, n. 18, p. 7332–7336, 1 maio 2007. Disponível em: <http://www.pnas.org/content/104/18/7332>. Acesso em: 6 maio 2016.

ONTOLOGY. In: GRUBER, Tom. (Ling Liu & M. Tamer Özsu, Org.)*Encyclopedia of Database Systems*. [S.l.]: Springer-Verlag, 2009. Disponível em: <http://tomgruber.org/writing/ontology-definition-2007.htm>. Acesso em: 31 mar. 2016.

OPEN DATA HANDBOOK DOCUMENTATION. *DIG: Decentralized Information Group*. [S.l.]: Open Knowledge Foundation, 2012. . Disponível em: <http://opendatahandbook.org/>. Acesso em: 22 fev. 2016.

OPEN DEFINITION. *Open Knowledge - Source Code*. 2.1. ed. [S.l: s.n.], 2015. . Disponível em: <http://opendefinition.org/>. Acesso em: 24 fev. 2016.

ORRANTIA, Josi Sierra. Conocity: videos enriquecidos con mapas para la gestión del conocimiento. In: FIFTH INTERNATIONAL CONFERENCE ON CONCEPT MAPPING, 17 set. 2012, Valleta, Malta. *Anais*... Valleta, Malta: University of Malta, IHMC, 17 set. 2012. Disponível em: <http://cmc.ihmc.us/cmc2012/CMC2012Program.html>. Acesso em: 22 fev. 2016.

PARENTE, André. *O virtual e o hipertextual*. Rio de Janeiro: Pazulin, 1999.

PAULHEIM, Heiko. Exploiting linked open data as background knowledge in data mining. *DMoLD*, v. 1082, 2013. Disponível em: <http://ceur-ws.org/Vol-1082/extendedAbstract.pdf>. Acesso em: 24 fev. 2016.

PEDROSO, Louise; TANAKA, Asterio; CAPPELLI, Claudia. A Lei de acesso à informação brasileira e os desafios tecnológicos dos dados abertos governamentais. In: IX SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO, 2013, João Pessoa. *Anais*... João Pessoa: UFPB, 2013. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/sbsi/2013/0048.pdf>. Acesso em: 25 fev. 2016.

PEREIRA, Frederico Cesar Mafra. A equação fundamental da Ciência da Informação e a importância de Brookes enquanto referência para o campo da Ciência da Informação. *Informação & Informação*, v. 13, n. 1, p. 15–31, 15 jul. 2008. Disponível em: <http://www.uel.br/revistas/wrevojs246/index.php/informacao/article/view/1761>. Acesso em: 30 mar. 2016.

PEREIRA, Júlio Cesar Rodrigues. *Análise de dados qualitativos: estratégias metodológicas para as ciências da saúde, humanas e sociais*. São Paulo: EDUSP : FAPESP, 2004.

PEZZI, Rafael Peretti. Ciência aberta: dos hipertextos aos hiperobjetos. In: ALBAGLI, SARITA; MACIEL, MARIA LUCIA; ABDO, ALEXANDRE HANNUD. *Ciência aberta, questões abertas*. Brasília, Rio de Janeiro: IBICTC, UNIRIO, 2015. p. 169–200. Disponível em: <http://livroaberto.ibict.br/handle/1/1060>.

PINHEIRO, Lena Vania Ribeiro. Pilares conceituais para mapeamento do território epistemológico da Ciência da Informação: disciplinaridade, interdisciplinaridade, transdisciplinaridade e aplicações. In: PINTO, VIRGINIA BENTES; CAVALCANTE, LÍGIA EUGÊNIA; SILVA NETO, CASEMIRO. [S.l.]: Edições UFC, 2007. p. 71–104. . Acesso em: 4 abr. 2016.

PONTES JUNIOR, João De; CARVALHO, Rodrigo de Aquino; AZEVEDO, Alexander Wilian. Da recuperação da informação à recuperação do conhecimento: reflexões e propostas. *Perspectivas em Ciência da Informação*, v. 18, n. 4, p. 02–17, 19 dez. 2013. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/965>. Acesso em: 29 fev. 2016.

PRESSMAN, Roger S. *Software Engineering: A Practitioner's Approach*. New York: McGraw-Hill, 2005.

RICHARDSON, Roberto Jarry. *Pesquisa social: métodos e técnicas*. 3. ed. rev e ampl ed. São Paulo: Atlas, 2012.

ROBINS, David. Informing Science Institute - Interactive Information Retrieval: Context and Basic Notions. *Informing Science*, v. 3, n. 2, p. 57–61, 2000. Disponível em: <http://inform.nu/Articles/Vol3/v3n2p57-62.pdf>. Acesso em: 10 abr. 2016.

ROBREDO, Jaime. *Da Ciência da Informação revisitada aos sistemas humanos de informação*. Brasília: Thesaurus e SSRR Informações, 2003a.

ROBREDO, Jaime. Epistemologia da ciência da informação revisitada. In: V ENCONTRO DA ASSOCIAÇÃO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO E BIBLIOTECONOMIA, 2003b, Belo Horizonte. *Anais*... Belo Horizonte: [s.n.], 2003. Disponível em: <http://repositorios.questoesemrede.uff.br/repositorios/handle/123456789/474>. Acesso em: 22 fev. 2016.

RODRIGUES, Maria Rosemary; CERVANTES, Brígida Maria Nogueira. Análise de assunto e mapas conceituais:semelhanças nos processos. *Perspectivas em Ciência da Informação*, v. 20, n. 4, p. 35–56, dez. 2015. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362015000400035&lng=pt&nrm=iso&tlng=en>. Acesso em: 18 mar. 2016.

RUSSELL, Bertrand. On propositions: what they are and how they mean. *Proceedings of the Aristotelian Society, Supplementary Volumes*, v. 2, p. 1–43, 1919. Disponível em: <http://www.jstor.org/stable/4106441>. Acesso em: 1 mar. 2016.

SALTON, Gerard; MCGILL, Michael J. *Introduction to Modern Information Retrieval*. USA: McGraw-Hill, 1983.

SANTOS NETO, Antonio Laurindo Dos *et al.* Tecnologias de dados abertos para interligar bibliotecas, arquivos e museus: um caso machadiano. *Transinformação*, v. 25, n. 1, p. 81–87, 2013. Disponível em: <http://www.scielo.br/pdf/tinf/v25n1/a08v25n1.pdf>. Acesso em: 30 maio 2016.

SARACEVIC, Tefko. Information science. *Journal of the American Society for Information Science*, v. 50, p. 1051–1063, 1999. Disponível em: <http://comminfo.rutgers.edu/~tefko/JASIS1999.pdf>.

SARACEVIC, Tefko. Information science. *Encyclopedia of Library and Information Sciences*. 3. ed. New York: Taylor & Francis, 2010. p. 2570–2586. Disponível em: <http://comminfo.rutgers.edu/~tefko/SaracevicInformationScienceELIS2009.pdf>. Acesso em: 16 fev. 2016.

SARACEVIC, Tefko. Interdisciplinary nature of information science. *Ciência da informação*, v. 24, n. 1, p. 36–41, 1995. Disponível em: <http://www.brapci.inf.br/_repositorio/2010/03/pdf_dd085d2c4b_0008887.pdf>. Acesso em: 22 fev. 2016.

SARACEVIC, Tefko. Relevance: a review of the literature and a framework for thinking on the notion in information science. Part III: behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, v. 58, n. 13, p. 2126–2144, 1 nov. 2007. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1002/asi.20681/full>. Acesso em: 16 fev. 2016.

SHADBOLT, N. *et al.* Web science: a new frontier. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, v. 371, n. 1987, p. 20120512–20120512, 18 fev. 2013. Disponível em: <http://rsta.royalsocietypublishing.org/cgi/doi/10.1098/rsta.2012.0512>. Acesso em: 19 fev. 2016.

SHAH, Rawn. The future of the social web depends on standards. *Forbes*, 2 jul. 2013. Disponível em: <http://www.forbes.com/sites/rawnshah/2013/07/02/the-future-of-the-social-web-depends-on-standards/>. Acesso em: 24 fev. 2016.

SHANNON, Claude Elwood. A mathematical theory of communication. *The Bell System Technical Journal*, v. 27, p. 379–423, 623–656, out. 1948. Acesso em: 23 fev. 2016.

SHERRATT, Christine S.; SCHLABACH, Martin L. The applications of concept mapping in reference and information services. *American Library Association*, v. 30, n. 1, p. 60–69, 1990. Disponível em: <http://www.jstor.org/stable/25828679>. Acesso em: 9 mar. 2016.

SILVA, Flávio Soares Corrêa Da; FINGER, Marcelo; MELO, Ana Cristina Vieira De. *Lógica para computação*. São Paulo: Thomson Learning, 2006.

SILVA, Renata Eleuterio Da; SANTOS, Plácida Leopoldina V. A. da Costa; FERNEDA, Edberto. Modelos de recuperação de informação e web semântica: a questão da relevância. *Informação & Informação*, v. 18, n. 3, p. 27, 9 out. 2013. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/12822>. Acesso em: 11 abr. 2016.

SILVA, Jonathas Carvalho; GOMES, Henriette Ferreira. Conceitos de informação na Ciência da Informação: percepções analíticas, proposições e categorizações. *Informação & Sociedade: Estudos*, v. 25, n. 1, p. 157, 5 fev. 2015. Disponível em: <http://www.ies.ufpb.br/ojs/index.php/ies/article/view/145>. Acesso em: 5 abr. 2016.

SOUSA, Paulo de Tarso Costa De. Metodologia de análise de redes sociais. In: MUELLER, SUZANA P. M. *Métodos para a pesquisa em Ciência da Informação*. Brasília: Tesaurus, 2007. p. 119.

SOUZA, Queila; QUANDT, Carlos. Metodologia de análise de redes sociais. In: DUARTE, F.; QUANDT, CARLOS; SOUZA, QUEILA. *O Tempo das redes*. São Paulo: [s.n.], 2008. p. 31–63. Disponível em: <https://www.academia.edu/257818/Metodologia_De_An%C3%A1lise_De_Redes_Sociais>. Acesso em: 31 mar. 2016.

SPINK, Amanda; COLE, Charles (Org.). *New directions in cognitive information retrieval*. Amsterdam: Springer Netherlands, 2005. v. 19. Disponível em: <http://link.springer.com/10.1007/1-4020-4014-8>. Acesso em: 29 jun. 2016. (The Information Retrieval Series).

STUCKENSCHMIDT, Heiner. Data semantics on the web. *Journal on Data Semantics*, Berlim, v. 1, n. 1, p. 1–9, 2012. Disponível em: <http://link.springer.com/article/10.1007/s13740-012-0003-z>. Acesso em: 17 fev. 2016.

STUCKENSCHMIDT, Heiner *et al.* On the Status of Experimental Research on the Semantic Web. *The Semantic Web–ISWC 2013*. [S.l.]: Springer, 2013. p. 591–606. Disponível em: <http://link.springer.com/chapter/10.1007/978-3-642-41335-3_37>. Acesso em: 22 fev. 2016.

STUCKENSCHMIDT, Heiner; NOESSNER, Jan; FALLAHI, Faraz. A Study in User-centric Data Integration. 2012, Setubal, Portugal. *Anais*... Setubal, Portugal: [s.n.], 2012. p. 5–14. Disponível em: <http://publications.wim.uni-mannheim.de/informatik/lski/noessner2012study.pdf>. Acesso em: 22 fev. 2016.

TERGAN, Sigmar-Olaf. Managing knowledge with computer-based mapping tools. 2003, Honolulu, HI, USA. *Anais*... Honolulu, HI, USA: University of Honolulu, 2003. p. 2514–2517. Disponível em: <https://www.editlib.org/p/14253/proceeding_14253.pdf>. Acesso em: 30 jun. 2016.

THAMMASUT, D.; SORNIL, O. A graph-based information retrieval system. In: 2006 INTERNATIONAL SYMPOSIUM ON COMMUNICATIONS AND INFORMATION TECHNOLOGIES, out. 2006, Ladkrabang, Thailand. *Anais*... Ladkrabang, Thailand: IEEE, out. 2006. p. 743–748. Disponível em: <http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=4141327>. Acesso em: 14 abr. 2016.

TODD, Ross J. Back to our beginnings: information utilization, Bertram Brookes and the fundamental equation of information science. *Information Processing & Management*, v. 35, n. 6, p. 851–870, nov. 1999. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0306457399000308>. Acesso em: 29 mar. 2016.

TRUONG, Quoc-Dinh *et al.* Information retrieval model based on graph comparison. In: JOURNÉES INTERNATIONALES D'ANALYSE STATISTIQUE DES DONNÉES TEXTUELLES (JADT), mar. 2008, Lyon, France. *Anais*... Lyon, France: Laboratoire ICAR - ENS-LSH, mar. 2008. Disponível em: <http://www.irit.fr/publis/SIG/2008_JADT_TDMC.pdf>. Acesso em: 14 abr. 2016.

USBECK, Ricardo. Combining linked data and statistical information retrieval: next generation information systems. *The semantic web: trends and challenges*. Cham, Switzerland: Springer International Publishing, 2014. p. 845–854. Disponível em: <http://mayor2.dia.fi.upm.es/oeg-upm/files/eswc2014/Phd%20Symposium/paper_6.pdf>.

VALERIO, Alejandro; LEAKE, David B.; CAÑAS, Alberto J. Using automatically generated concept maps for document understanding: a human subjects experiment. In: FIFTH INTERNATIONAL CONFERENCE ON CONCEPT MAPPING, 17 set. 2012, Valleta, Malta. *Anais...* Valleta, Malta: University of Malta, IHMC, 17 set. 2012. Disponível em: <http://cmc.ihmc.us/cmc/CMCProceedings.html>. Acesso em: 25 fev. 2016.

VAN RIJSBERGEN, Cornelis Joost. *Information retrieval*. London: Butterworths, 1979. Disponível em: <http://openlib.org/home/krichel/courses/lis618/readings/rijsbergen79_infor_retriev.pdf>. Acesso em: 19 abr. 2016.

VEKIRI, Ioanna. What is the value of graphical displays in learning? *Educational Psychology Review*, v. 14, n. 3, p. 261–312, 2002. Disponível em: <http://link.springer.com/article/10.1023/A:1016064429161>. Acesso em: 21 fev. 2016.

VICKERY, Alina; VICKERY, Brian C. *Information science in theory and practice*. London: Butterworth-Heinemann, 1987.

W3C - Ontologies. Web page: MIT, ERCIM, Keio, Beihang, 2015. Disponível em: <https://www.w3.org/standards/semanticweb/ontology>. Acesso em: 4 maio 2016.

WARE, Colin. *Visual thinking for design*. [S.l.]: ELSEVIER, Morgan Kaufmann, 2010. . Acesso em: 11 mar. 2016.

WASSERMAN, Stanley; FAUST, Katherine. *Social network analysis: methods and applications*. Cambridge, England; New York: Cambridge University Press, 1994.

WATTS, Duncan J.; STROGATZ, Steven H. Collective dynamics of "small-world" networks. *Nature*, v. 393, n. 6684, p. 440–442, 4 jun. 1998. Disponível em: <http://www.nature.com/nature/journal/v393/n6684/abs/393440a0.html>. Acesso em: 24 fev. 2016.

WERSIG, Gernot; NEVELING, Ulrich. The phenomena of interest to information science. *The information scientist*, v. 9, n. 4, p. 127–140, 1975. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.232.5319&rep=rep1&type=pdf>. Acesso em: 16 fev. 2016.

WILLIAMS, Carol G. Using Concept Maps to Assess Conceptual Knowledge of Function. *Journal for Research in Mathematics Education*, v. 29, n. 4, p. 414–421, 1998. Disponível em: <http://www.jstor.org/stable/749858>. Acesso em: 4 mar. 2016.

WILSON, Thomas D. Human information behavior. *Informing science*, v. 3, n. 2, p. 49–56, 2000. Disponível em: <http://210.48.147.73/silibus/human.pdf>. Acesso em: 22 fev. 2016.

WILSON, Thomas D. Models in information behaviour research. *Journal of Documentation*, v. 55, n. 3, p. 249–271, 1999. Disponível em: <http://www.informationr.net/tdw/publ/papers/1999JDoc.html>. Acesso em: 24 fev. 2016.

YAO, Yiyu *et al.* Knowledge retrieval (kr). 2007, [S.l.]: IEEE, 2007. p. 729–735. Disponível em: <http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4427181>. Acesso em: 19 abr. 2016.

ZAVERI, Amrapali *et al.* Quality assessment for linked data: A survey. *Semantic Web*, v. 7, n. 1, p. 63–93, 2015. Disponível em: <http://content.iospress.com/articles/semantic-web/sw175>. Acesso em: 22 fev. 2016.

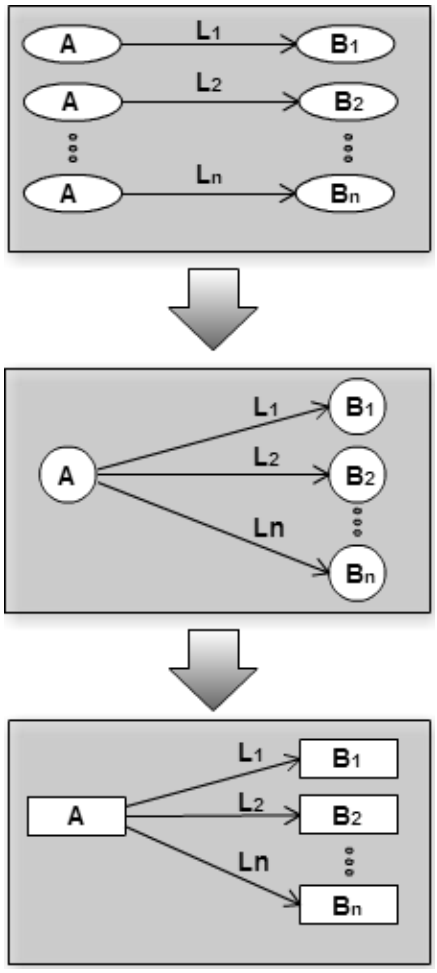
ZHANG, Jin. *Visualization for information retrieval*. Berlin: Springer, 2008. (The information retrieval series).

ZINS, Chaim. Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology*, v. 58, n. 4, p. 479–493, 15 fev. 2007. Disponível em: <http://doi.wiley.com/10.1002/asi.20508>. Acesso em: 23 fev. 2016.

APÊNDICE A – Mapeamentos entre RDFs, rede de informação e mapa conceitual

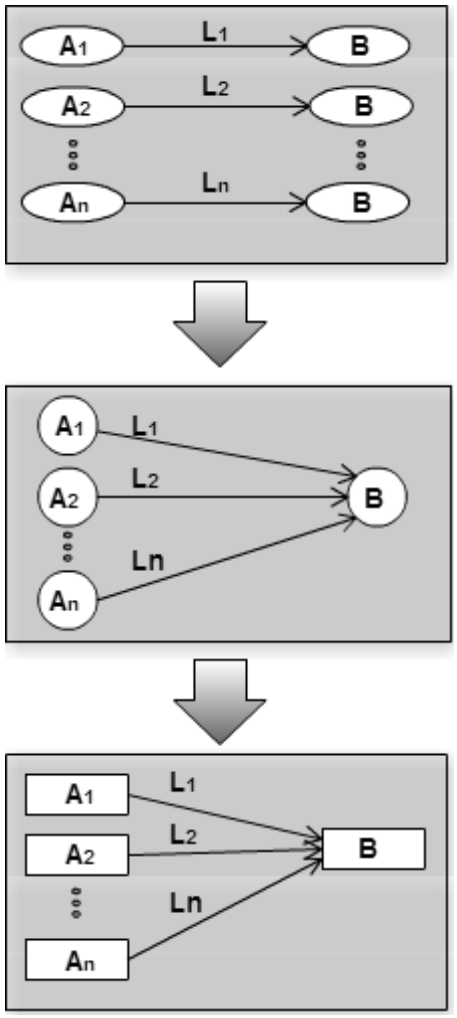
A Figura 43 e a Figura 101 dão continuidade aos casos de mapeamento descritos na subseção 4.3.1, apresentando as outras situações de igualdade ou diferença entre os sujeitos, predicados e objetos. Duas situações que merecem explicação adicional: a Figura 97 possui atributos na rede para representar existência de múltiplas ligações na mesma conexão, devido à limitação da rede em aceitar mais de uma conexão entre dois nós; a Figura 101 transforma um predicado catalogado (aqueles escolhidos por terem informações especiais, tal como o resumo associado a um sujeito) num atributo da conexão da rede e, posteriormente, num *hint* (uma janela informativa que aparece o ponteiro do mouse passa sobre o conceito).

Figura 91 – Sujeitos iguais, predicados e objetos diferentes



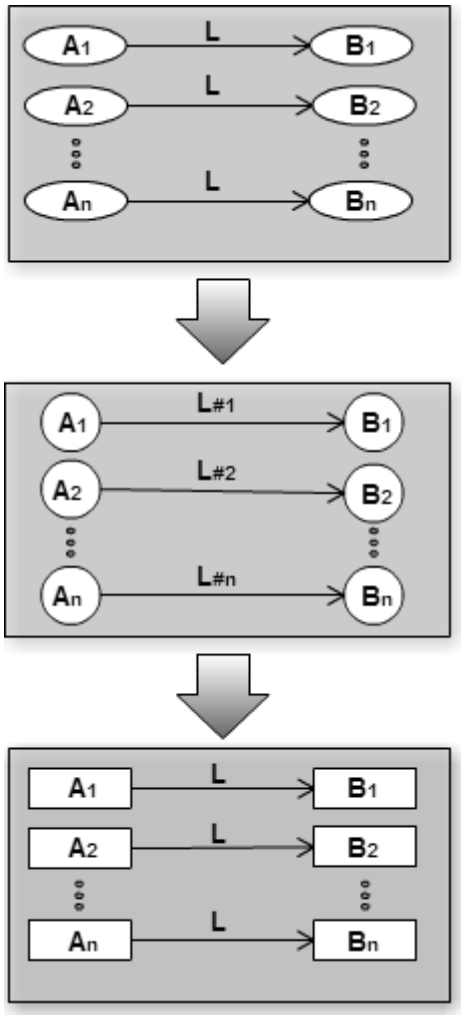
Fonte: Elaboração própria

Figura 92 – Sujeitos e predicados diferentes, objetos iguais



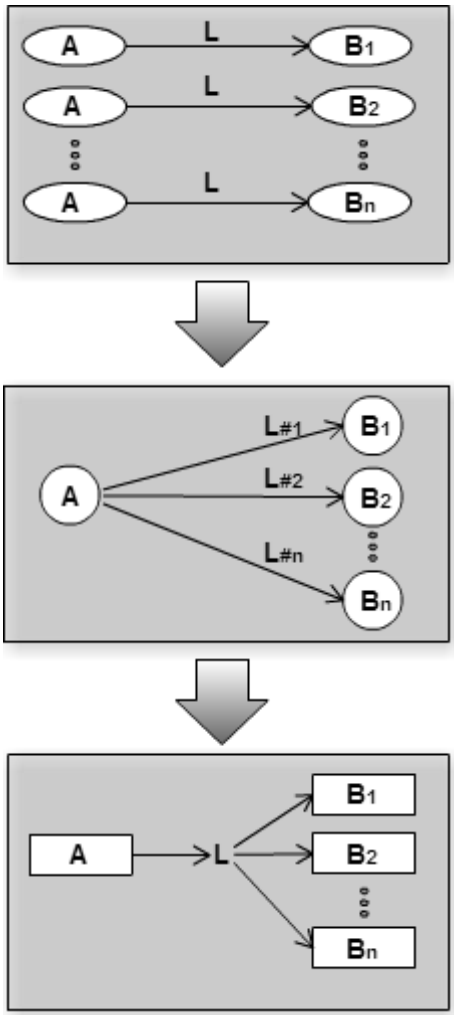
Fonte: Elaboração própria

Figura 93 – Predicados iguais, sujeitos e objetos diferentes



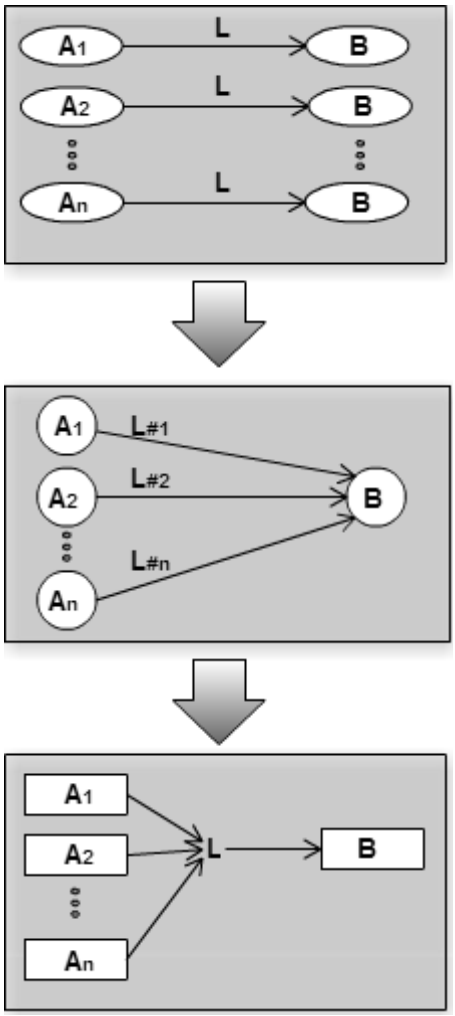
Fonte: Elaboração própria

Figura 94 – Sujeitos e predicados iguais, objetos diferentes



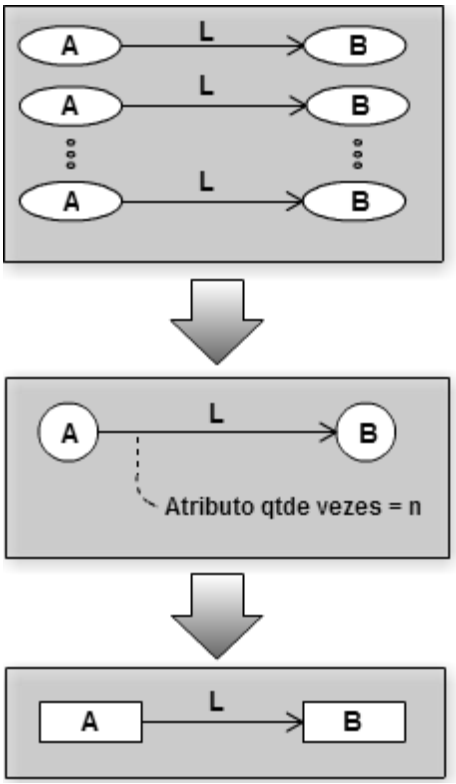
Fonte: Elaboração própria

Figura 95 – Objetos e predicados iguais, sujeitos diferentes



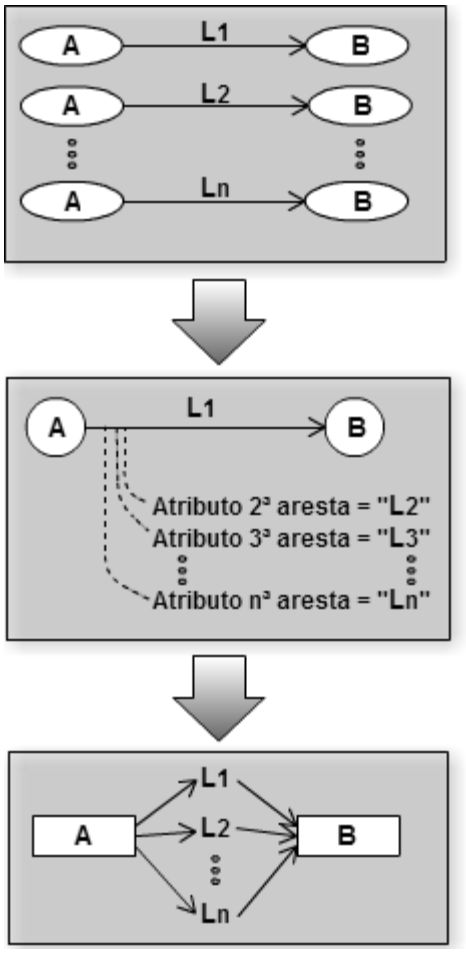
Fonte: Elaboração própria

Figura 96 – Sujeitos, predicados e objetos iguais



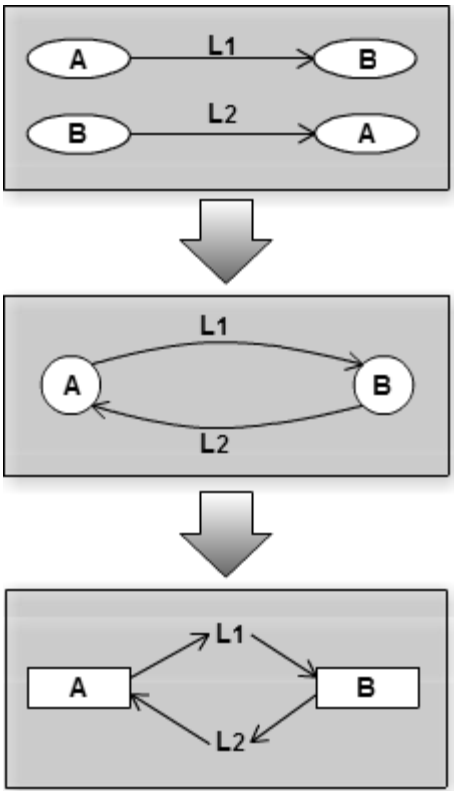
Fonte: Elaboração própria

Figura 97 – Sujeitos e objetos iguais, predicados diferentes



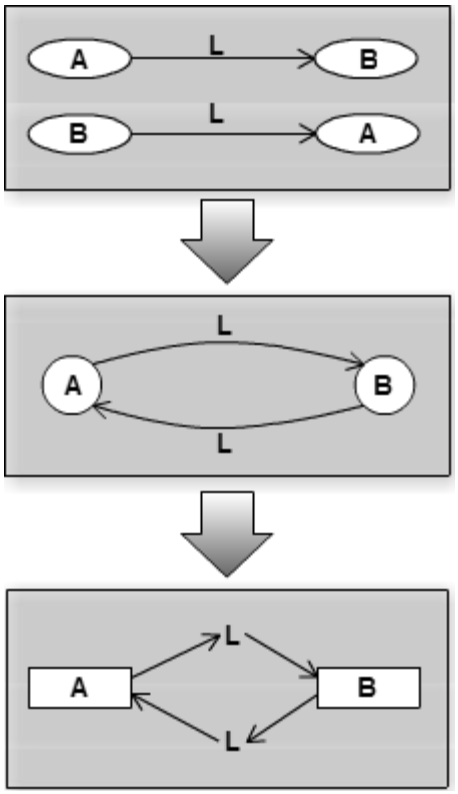
Fonte: Elaboração própria

Figura 98 – Sujeitos e objetos invertido, predicados diferentes



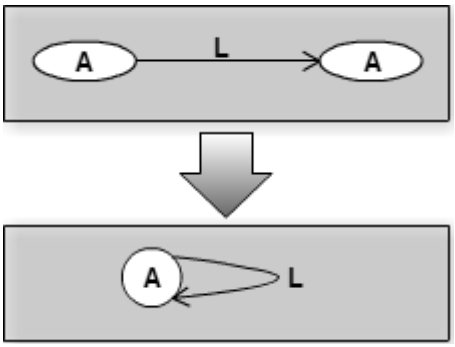
Fonte: Elaboração própria

Figura 99 – Sujeitos e objetos invertidos, predicados iguais



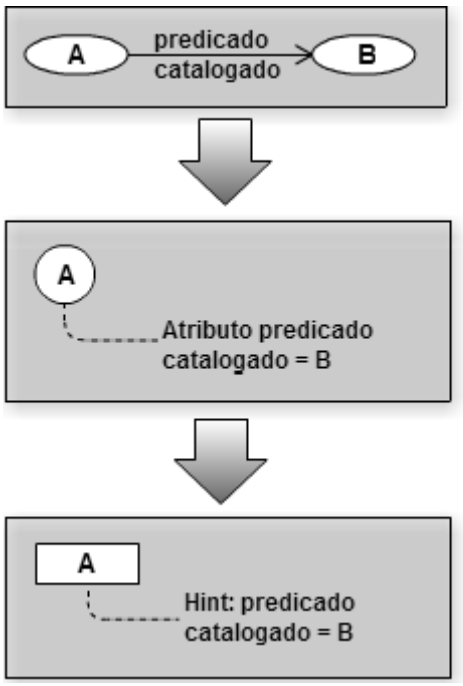
Fonte: Elaboração própria

Figura 100 – Sujeito e objeto iguais entre si



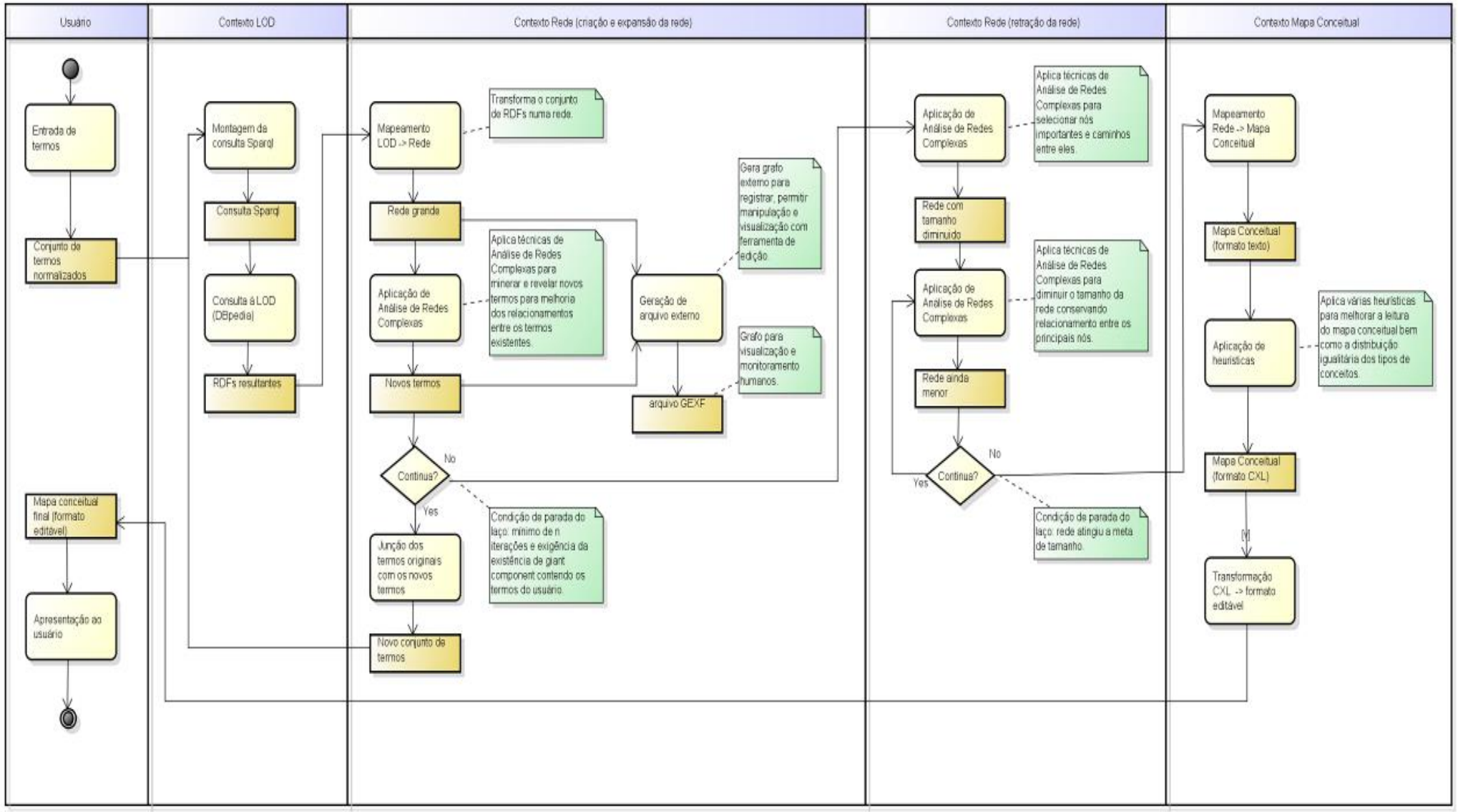
Fonte: Elaboração própria

Figura 101 – Predicado é catalogado

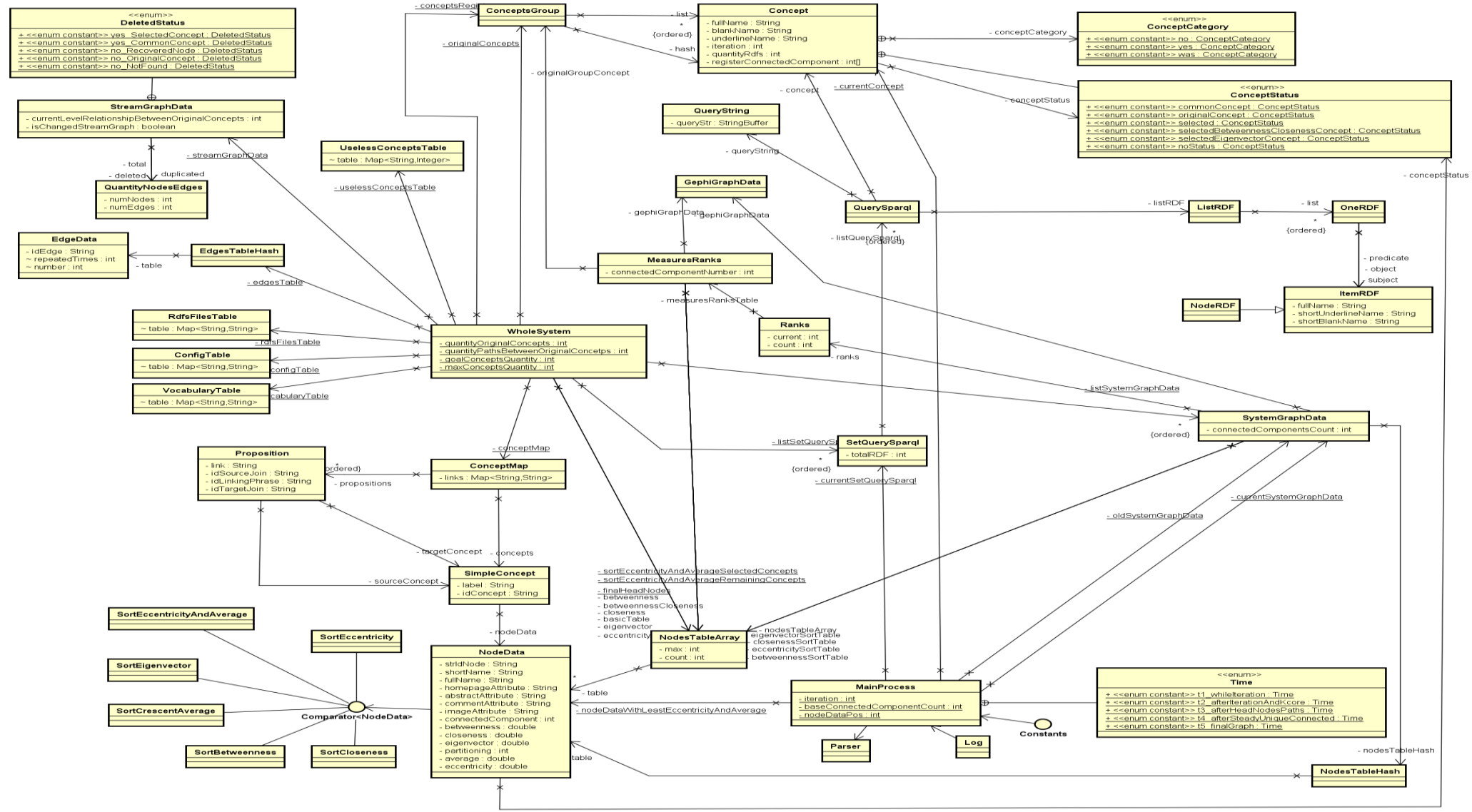


Fonte: Elaboração própria

APÊNDICE B – Diagrama de atividades detalhado do modelo aprimorado



APÊNDICE C – Diagrama de classes do protótipo



APÊNDICE D – Consulta modelo para a DBpedia

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX : <http://dbpedia.org/resource/>
PREFIX dbpedia: <http://dbpedia.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>

PREFIX relationship: <http://relationship/>

CONSTRUCT {
   <http://dbpedia.org/resource/#######> ?predicate ?object .
   ?subject ?predicate2 <http://dbpedia.org/resource/#######> .
   <http://dbpedia.org/resource/#######> relationship:homepage ?homepage .
   <http://dbpedia.org/resource/#######> relationship:comment ?comment .
   <http://dbpedia.org/resource/#######> relationship:abstract ?abstract .
   <http://dbpedia.org/resource/#######> relationship:image ?image .
}
WHERE {
   {
      <http://dbpedia.org/resource/#######> ?predicate ?object .
      FILTER regex(?object, "http://dbpedia.org/resource/")
   }
   UNION
   {
      ?subject ?predicate2 <http://dbpedia.org/resource/#######> .
      FILTER regex(?subject, "http://dbpedia.org/resource/")
   }
   UNION
   {
      <http://dbpedia.org/resource/#######> foaf:isPrimaryTopicOf ?homepage
.
   }
   UNION
   {
      <http://dbpedia.org/resource/#######> rdfs:comment ?comment .
      FILTER(lang(?comment)="en")
   }
   UNION
   {
      <http://dbpedia.org/resource/#######>
<http://dbpedia.org/ontology/abstract> ?abstract .
      FILTER(lang(?abstract)="en")
   }
   UNION
   {
      <http://dbpedia.org/resource/#######>
<http://dbpedia.org/property/image> ?image .
      FILTER(lang(?image)="en")
   }
}
```

APÊNDICE E – Configuração do protótipo

```
//============================================================================
// BASIC CONFIG
//============================================================================

// Current test name
testName = test01

// Knowledge base local (1 = DBpedia; 2 = my knowledge base, in this case
directedStreamGraph was changed to true)
knowledgeBasePlace = 2

// Base directory to creation of output files: absolut path
'C:\\path\\results' or relative path 'path\\results' or current path '.'
// baseDirectory = C:\\Users\\Henrique\\Documents\\data_collect
baseDirectory = C:\\documentos\\Doutorado\\Tese\\my_knowledge_base

// minimum quantity of iterations to get out of loop when connected
component = 1 (good value: 6)
minIterationToVerifyUniqueConnectedComponent = 6

// minimum quantity of iterations to get out of loop when exist
relationship between original concepts (good value: 10)
minIterationToVerifyRelationshipBetweenOriginalConcepts = 8

// maximum quantity of iterations to get out of loop (good value: 50)
maxIteration = 50

// N-degree filter used in all system (good value: 2)
nDegreeFilter = 2

// apply n-degree filter trigger from iteration number (good value: 2)
iterationTriggerApplyNDegreeFilterAlgorithm = 3

// quantity of nodes to shoot n-degree filter algorithm (good value: 10000)
quantityNodesToApplyNdegreeFilter = 10000

// apply n-degree filter algorithm only if connected component = 1 (good
value: true)
isUniqueConnectedComponentToApplyNdegreeFilter = false

// make the duplication of concept: from with "category" to without one
(change to modify quantity of concepts with "category")
additionNewConceptWithoutCategory = true

// K-core used in all system (good value: 2)
kCoreFilter = 2

// quantity of nodes to shoot K-core n algorithm
quantityNodesToApplyKcoreFilter = 700

// fator added in logarithmic function to calculate goal of concepts
quantity in final concept to map (good value: 3)
// function: log2(1/countOriginal)+4+factor
conceptsQuantityCalculationFactor = 0

// top limit to range of concepts quantity (good value: 5)
conceptsMinMaxRange = 3
```

```
//============================================================================
// ADVANCED CONFIG
//============================================================================

// Name of file with the concepts (it is recommended an amount of 2 to 12
concepts)
nameUserTermsFile = terms.txt

// use of the useless table to discart terms in rdfs collect fase
isEnableUselessTable = true

// BETWEENNESS + CLOSENESS
// quantity of nodes to selection in all connected componentes (proporcion
relative to quantity total of user terms) (good value: 3.0)
proporcionBetweennessCloseness = 1.0
// precision added up to rounding the calculate of quantity of each
connected component (good value: 0.5)
precisionBetweennessCloseness = 0.5
// quantity of nodes to use as base to build betweenness+closeness sorted
table.
// proporcion relative to quantity total of user terms. Must be >=
maxBetweennessCloseness. (good value: double of maxBetweennessCloseness)
// (if near maxBetweennessCloseness than the sort will disregard closeness)
proporcionBetweennessOnly = 10
// maximum limit to quantity of new concepts distributed in all connected
component, +0.5 in each component (excluding the addition of nodes by
"Category:")
// proporcion relative to quantity total of user terms. (good value: 5)
maxBetweennessCloseness = 5

// EIGENVECTOR
// quantity of nodes to selection in all connected componentes (proporcion
relative to quantity total of original nodes) (good value: 1.0)
proporcionEigenvector = 1.0
// precision added up to rounding the calculate of quantity of each
connected component (good value: 0.5)
precisionEigenvector = 0.5
// maximum limit to quantity of new concepts distributed in all connected
component, +0.5 in each component (excluding the addition of nodes by
"Category:")
// proporcion relative to quantity total of user terms. (good value: 15)
maxEigenvector = 5

// choice of nodes to be head. They are used to build the shortest paths
(obs. the original concepts are always chosen) (good value:
isBetweennessCloseness = true)
isBetweennessCloseness = true
isEigenvector = false
isSelected = false

// keep all nodes with link to original concepts (in stage after selection
of head nodes) (good value: false)
// (normally this flag improves much more the final quantity of concepts in
concept map)
isKeepNeighborsOfOriginalConcepts = false

// quantity of nodes each rank to show in short report (good value: 0)
quantityNodesShortReport = 0

// color of original concepts (good value: 200.200.200.255)
```

```
backGroundcolorOriginalConcept = 200.200.200.255

// thickness of concepts with hint (good value: 2)
borderThicknessConceptWithHint = 2

// line maximum length of sentences in the final concept map
maxLineLengthConcept    = 15
maxLineLengthLinkPhrase = 15

// Names of files  (## = will be change to value of the testName var)
nameQueryDefaultFile    = query_model\\query.txt
nameVocabularyFile      = vocabulary\\linkvocabulary.txt
nameUselessConceptsFile = vocabulary\\uselessconcepts.txt
nameTxtConceptMapFile   = conceptmap_##.txt
nameCxlConceptMapFile   = conceptmap_##.cxl
nameGexfGraphFile       = graph_##.gexf
nameCompleteReportFile  = complete_report_##.txt
nameShortReportFile     = short_report_##.txt
nameConsoleReportFile   = console_report_##.txt
nameConsoleErrorFile    = consoleErr_##.txt
nameMyKnowledgeBaseFile = myKnowledgeBase\\base.txt

// Names of directories
dirRdfsPersistenceFiles = persistenceRdfs
dirGraph                = graph
dirLog                          = log
dirConceptMap           = conceptMap

// DBPEDIA server (good value: http://dbpedia.org/sparql) (alternative good
value: http://lod.openlinksw.com/sparql)
dbpediaServer = http://dbpedia.org/sparql

// to see simultaneously the build of graph in Gephi (good value: false)

gephiVisualization = false

// to see simultaneously the build of graph in Graph Stream Visualization
(good value: false)
graphStreamVisualization =   false

// to fix a bug in Gephi Tool Kit - calculate wrong the quantity of
connected component
// turn the flag to true and put original concepts that is getting alone
(put '.' to empty)
isFixBugInGephiToolKit = false
originalConceptWithGephiToolKitBug1 = .
originalConceptWithGephiToolKitBug2 = .
originalConceptWithGephiToolKitBug3 = .
```

APÊNDICE F – Vocabulário controlado usado no protótipo

```
almaMater -> has alma matter
author -> is author
birthPlace -> birth place
citizenship -> has citizenship in
core#broader -> is broader of
core#related -> is related to
core#subject -> core subject
deathPlace -> death place
designer -> was designed for
developer -> was developed for
doctoralAdvisor -> had as doctoral advisor
doctoralStudent -> had as doctoral student
doctoralStudents -> had as doctoral student
east -> is in the east
field -> works in the field
fields -> works in the fields
focus -> has focus in
foundationPlace -> has place foundation
genre -> has genre
industry -> is a industry of
influencedBy -> influenced by
influences -> received influences of
isPartOf -> is part of
knownFor -> is known for
leaderName  -> is leader
license -> has license
location -> has location in
locationCountry -> is located in
mainInterest -> has as main interest
mainInterests -> has as main interest
nationality -> has the nationality
nonFictionSubject -> is non-fiction subject of
north -> is in the north
northeast -> is in the northeast
northwest -> is in the northwest
notableIdea -> notable idea
occupation -> has occupacion
owl#differentFrom -> is different of
owner -> is owned by
placeOfBirth -> birth place
rdf-schema#seeAlso -> see also
residence -> has residence in
schoolTradition -> has school tradition
shortDescription -> has short description
skills -> has skills
south -> is in the south
southeast -> is in the southeast
southwest -> is in the southwest
subdivisionType -> is a subdivision of
subject     -> has subject
type -> is a type of
west -> is in the west
wikiPageDisambiguates -> maybe
wikiPageRedirects -> see also
```

APÊNDICE G – Formulário de coleta de dados sobre a avaliação do sistema

Nesse apêndice encontram-se as três páginas do formulário citado. O método de uso está explicado no capítulo da metodologia seção 3.5, e os dados coletados são apresentados na seção 4.6 e discutidos nas seções 5.6.1 e 5.6.2.

PPGCINF/UnB
Pesquisa: Uma proposta de Metodologia para Recuperação de Informação no contexto de Dados Ligados
Doutorando: Henrique Monteiro Cristovão
Ano: 2015

Coleta de dados (2ª etapa)

Nessa etapa você avaliará três mapas conceituais resultantes construídos automaticamente pelo sistema. Os dois primeiros tiveram como ponto de partida os dois conjuntos de termos que você enviou na 1ª etapa da coleta de dados. O terceiro foi feito a partir dos termos fornecidos por outro usuário.

Instruções:

Abra os três mapas conceituais anexos ao email que lhe foi enviado (mapa1, mapa2 e mapa3). Eles estão nos formatos PDF e CMAP. Abra-os no formato PDF, porém, caso você tenha o software CmapTools instalado em seu computador prefira abrir os mapas no formato CMAP, pois será possível ver informações adicionais passando o ponteiro do mouse sobre alguns conceitos identificados com borda mais grossa. No entanto, a falta do CmapTools não inviabiliza o experimento, pois as questões desse questionário não fazem referência a essas informações adicionais associadas aos conceitos. De forma opcional , você pode instalar o CmapTools - http://cmap.ihmc.us/.

Preencha o formulário que segue.

1. Seu nome:
2. Qual seu nível de conhecimento na leitura de mapas conceituais: ○ Bom, ○ Pouco, ○ Nenhum.
3. Qual seu nível de conhecimento na criação de mapas conceituais: ○ Bom, ○ Pouco, ○ Nenhum.
4. Considerando que o principal objetivo do sistema proposto é a descoberta automática de relações entre os termos fornecidos pelo usuário, seja por meio de ligações diretas ou ligações indiretas usando novos conceitos. Considerando também que não é foco do sistema lidar de forma individual com os conceitos, seja dando explicações, detalhando ou trazendo características e informações individuais de cada um. Responda as questões da página seguinte para cada um dos três mapas resultantes em anexo.

   Obs.: os termos fornecidos por você ou pelo usuário (aqui chamados de termos base) estão destacados em caixas cinza.

Questões	Mapa 1 (três termos base fornecidos por você)	Mapa 2 (seis termos base fornecidos por você)	Mapa 3 (três termos base fornecidos por outro usuário)
a) Qual seu conhecimento dos conceitos abordados pelo mapa conceitual?	◯ conheço o suficiente para poder avaliar. ◯ Não conheço o suficiente. (neste caso, não responda os itens restantes de 'b' a 'g')	◯ conheço o suficiente para poder avaliar. ◯ Não conheço o suficiente. (neste caso, não responda os itens restantes de 'b' a 'g')	◯ conheço o suficiente para poder avaliar. ◯ Não conheço o suficiente. (neste caso, não responda os itens restantes de 'b' a 'g')
b) Quanto esse mapa conceitual lhe ajudaria a entender as relações entre os termos base?	◯ ajudaria muito ◯ ajudaria razoavelmente ◯ ajudaria pouco ◯ não ajudaria	◯ ajudaria muito ◯ ajudaria razoavelmente ◯ ajudaria pouco ◯ não ajudaria	◯ ajudaria muito ◯ ajudaria razoavelmente ◯ ajudaria pouco ◯ não ajudaria
c) Quanto esse mapa conceitual lhe serviria como ponto de partida para auxiliar uma pesquisa sobre as relações entre os termos base?	◯ ajudaria muito ◯ ajudaria razoavelmente ◯ ajudaria pouco ◯ não ajudaria	◯ ajudaria muito ◯ ajudaria razoavelmente ◯ ajudaria pouco ◯ não ajudaria	◯ ajudaria muito ◯ ajudaria razoavelmente ◯ ajudaria pouco ◯ não ajudaria
d) Quanto esse mapa lhe ajudaria a iniciar a construção de um mapa conceitual a partir do conjunto de termos base tendo com o principal objetivo estabelecer relações entre eles?	◯ ajudaria muito ◯ ajudaria razoavelmente ◯ ajudaria pouco ◯ não ajudaria	◯ ajudaria muito ◯ ajudaria razoavelmente ◯ ajudaria pouco ◯ não ajudaria	◯ ajudaria muito ◯ ajudaria razoavelmente ◯ ajudaria pouco ◯ não ajudaria
e) A respeito da relevância dos novos conceitos trazidos pelo sistema, quantos deles você acha que podem contribuir para o entendimento das relações entre os termos base?	Percentual aproximado 0 a 100%:	Percentual aproximado 0 a 100%:	Percentual aproximado 0 a 100%:

Questões (continuação)	Mapa 1 (três termos base fornecidos por você)	Mapa 2 (seis termos base fornecidos por você)	Mapa 3 (três termos base fornecidos por outro usuário)
f) A respeito da relevância das proposições (conceito -> ligação -> conceito), quantas delas você acha que podem contribuir para o entendimento das relações entre os termos base?	Percentual aproximado 0 a 100%:	Percentual aproximado 0 a 100%:	Percentual aproximado 0 a 100%:
g) Quantas proposições você julga que faltaram nesse mapa, considerando-o como primeira versão? Entenda-se por 'faltaram' proposições que você considera primordiais para estabelecer relacionamentos entre os termos base. Cite as mais importantes. Escreva: conceito, ligação, conceito.	Quantidade aproximada:	Quantidade aproximada:	Quantidade aproximada:

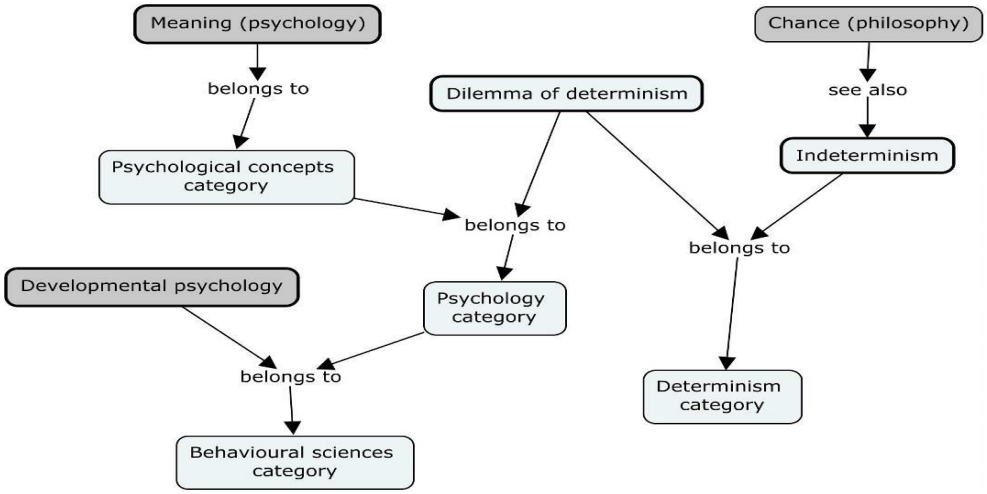
Quando terminar, salve este PDF preenchido e retorne-o anexado ao email de resposta.

Obrigado pela sua contribuição.

APÊNDICE H – Mapas conceituais resultantes da coleta de dados com os usuários

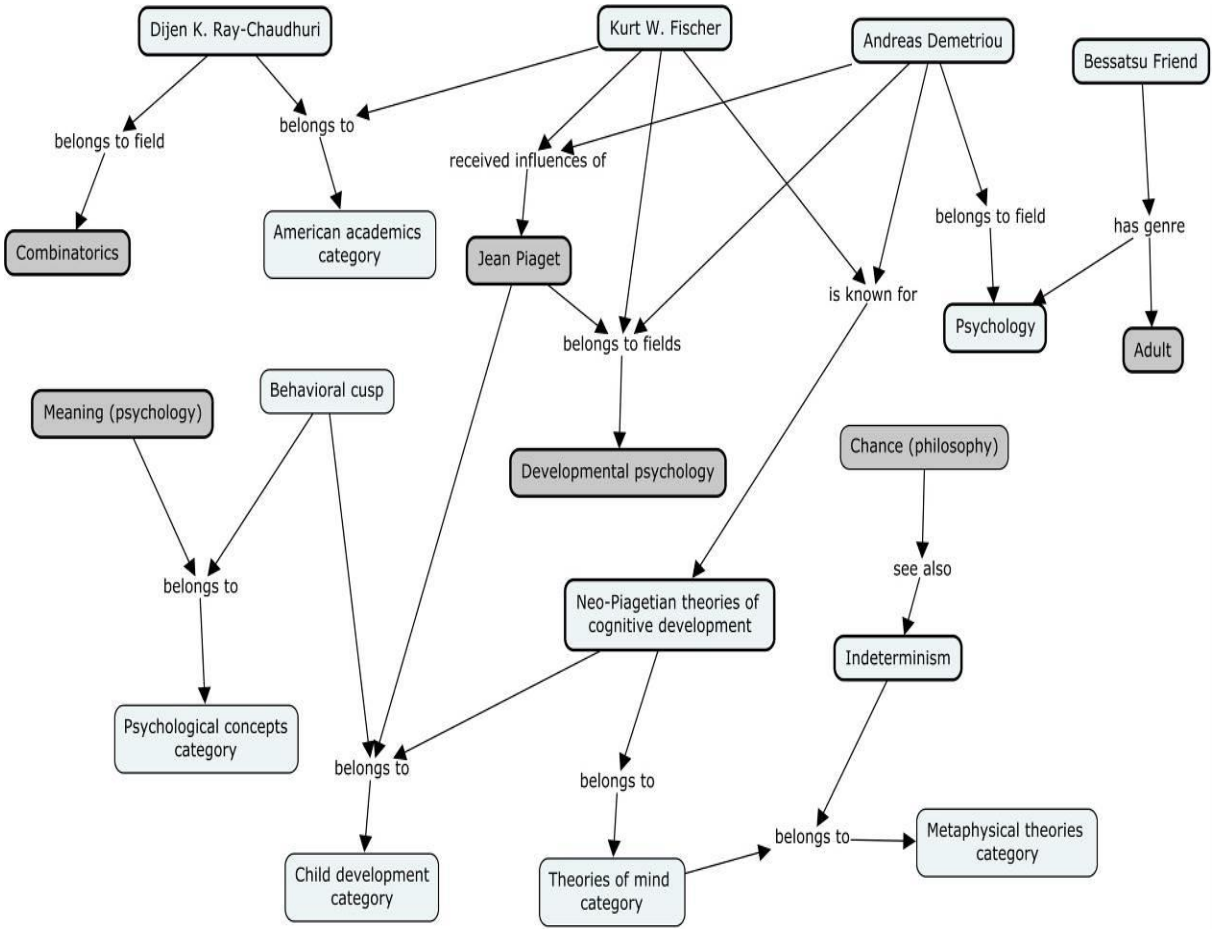
Cada usuário que participou da coleta de dados, explicada na seção 3.5, cujos dados são apresentados na seção 4.6 e discutidos nas seções 5.6.1 e 5.6.2, forneceu dois conjuntos de termos, o primeiro com três termos e o segundo com seis termos. Por intermédio do protótipo apresentado na seção 4.5, foi gerado automaticamente, para cada conjunto de termos, um mapa conceitual resultante onde os termos do usuário estão destacados com caixa de fundo cinza. A Figura 102 até a Figura 133 mostram esses mapas conceituais resultantes.

Figura 102 – Mapa resultante do usuário 1, a partir de três termos



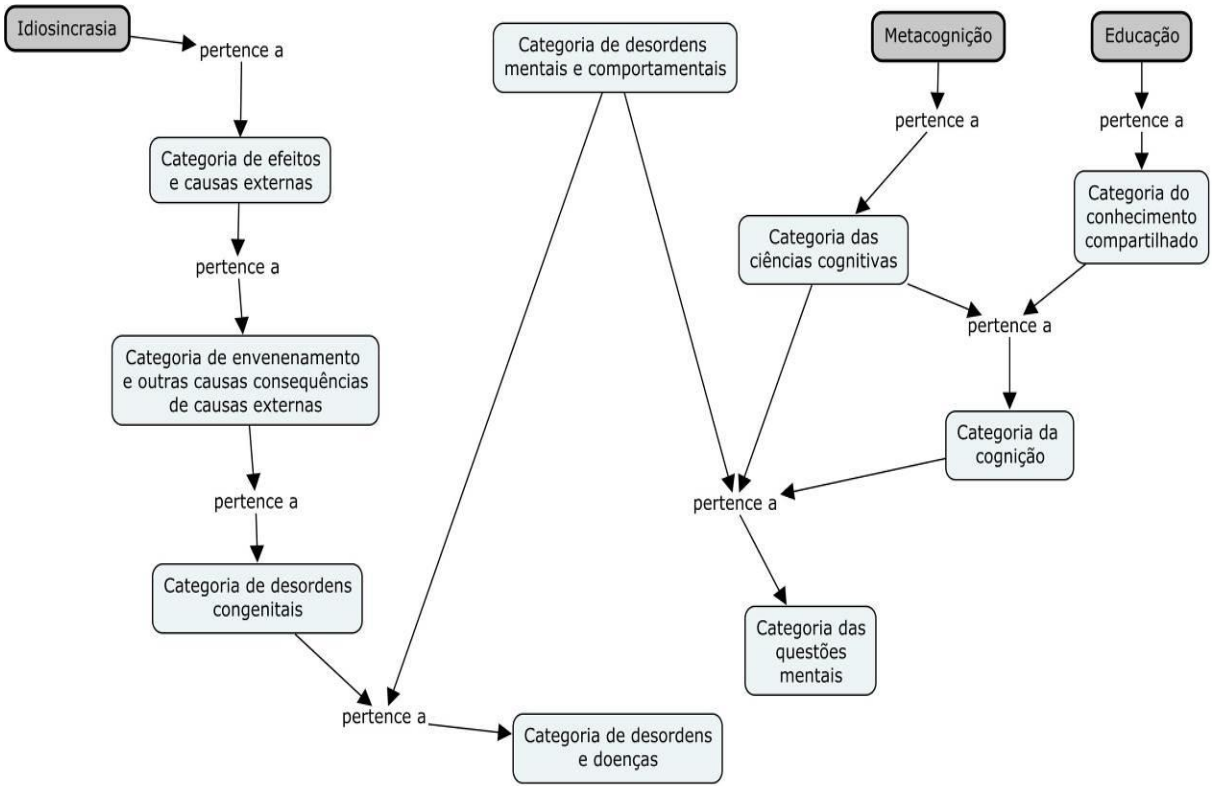
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 103 – Mapa resultante do usuário 1, a partir de 6 termos



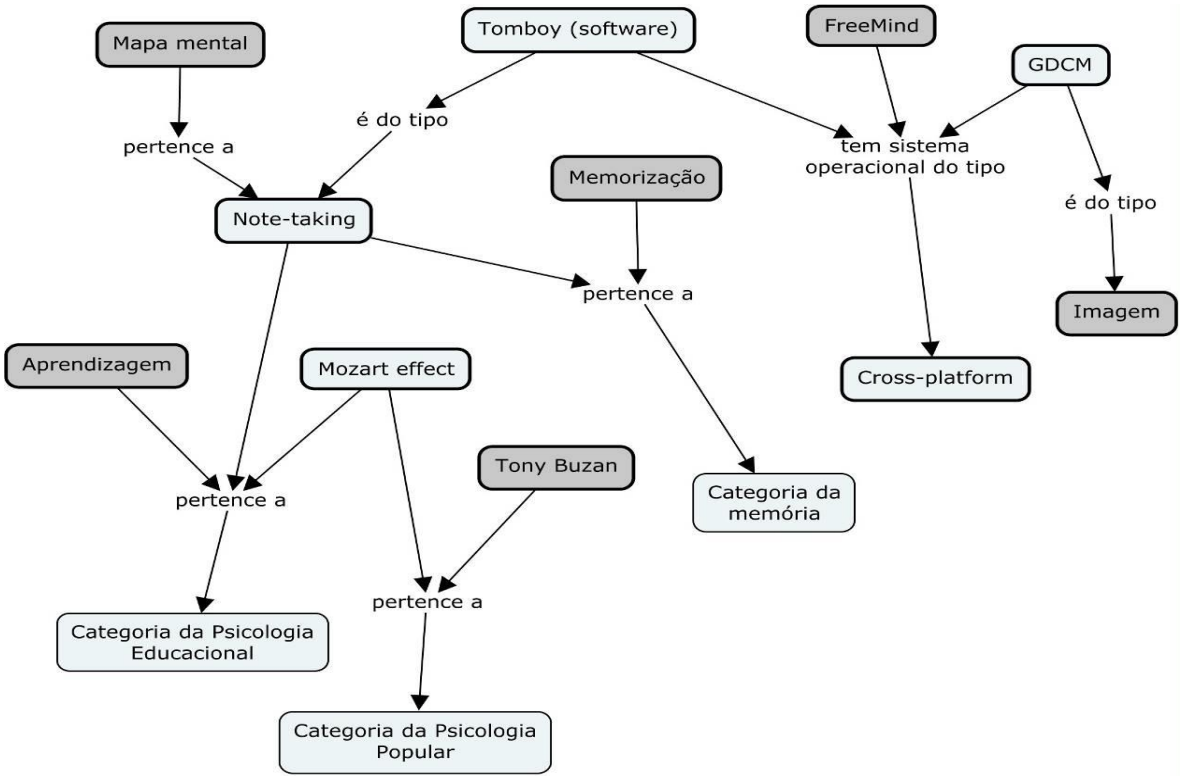
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 104 – Mapa resultante do usuário 2, a partir de três termos



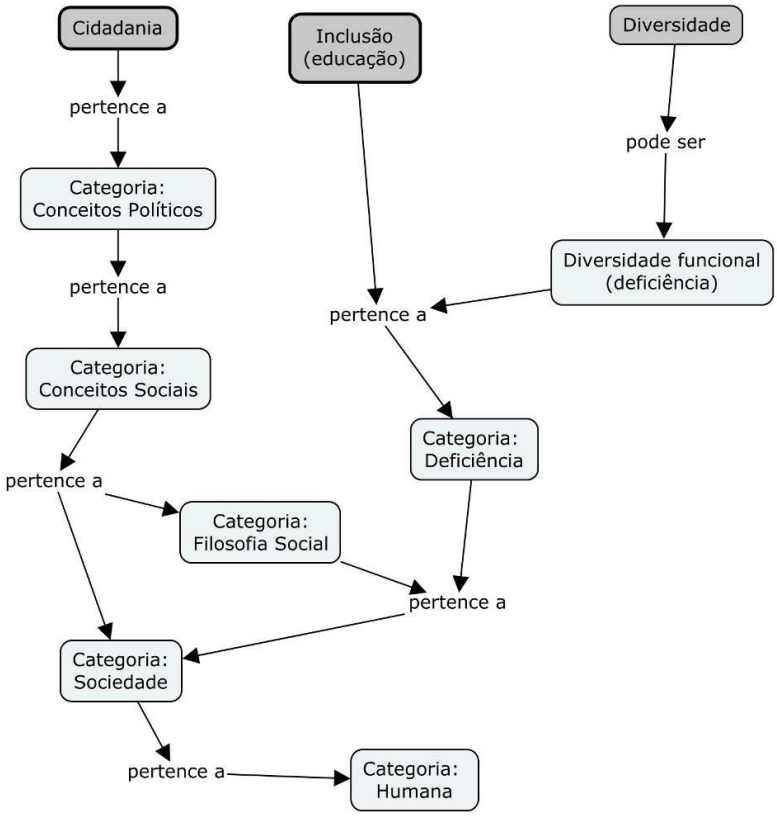
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 105 – Mapa resultante do usuário 2, a partir de seis termos



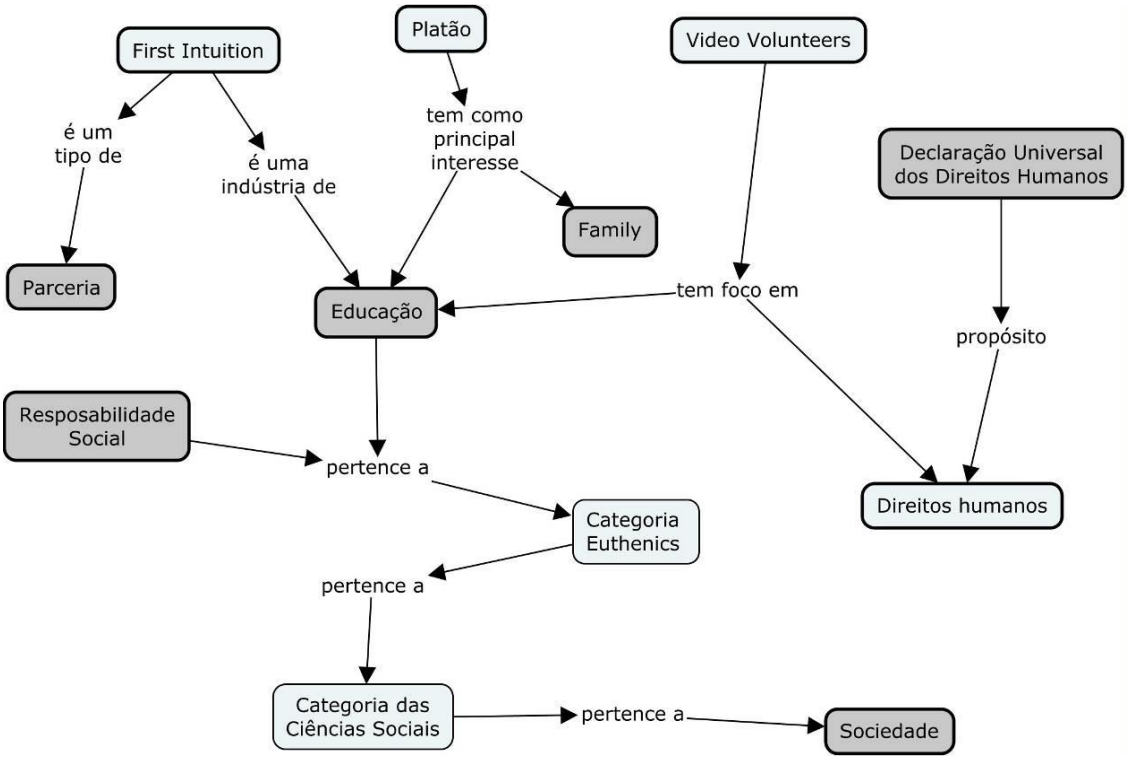
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 106 – Mapa resultante do usuário 4, a partir de três termos



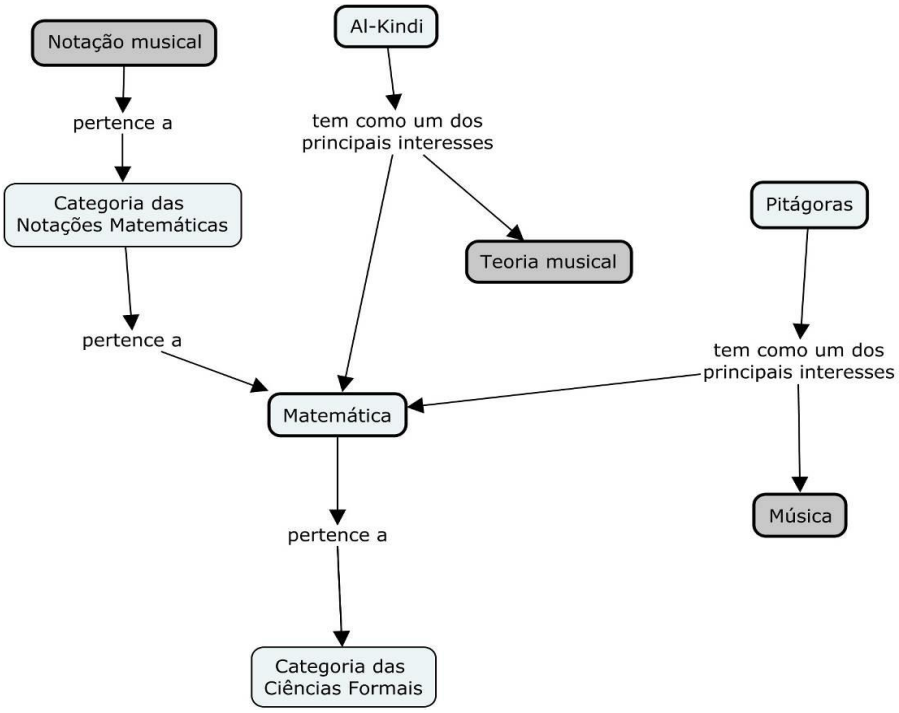
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 107 – Mapa resultante do usuário 4, a partir de seis termos



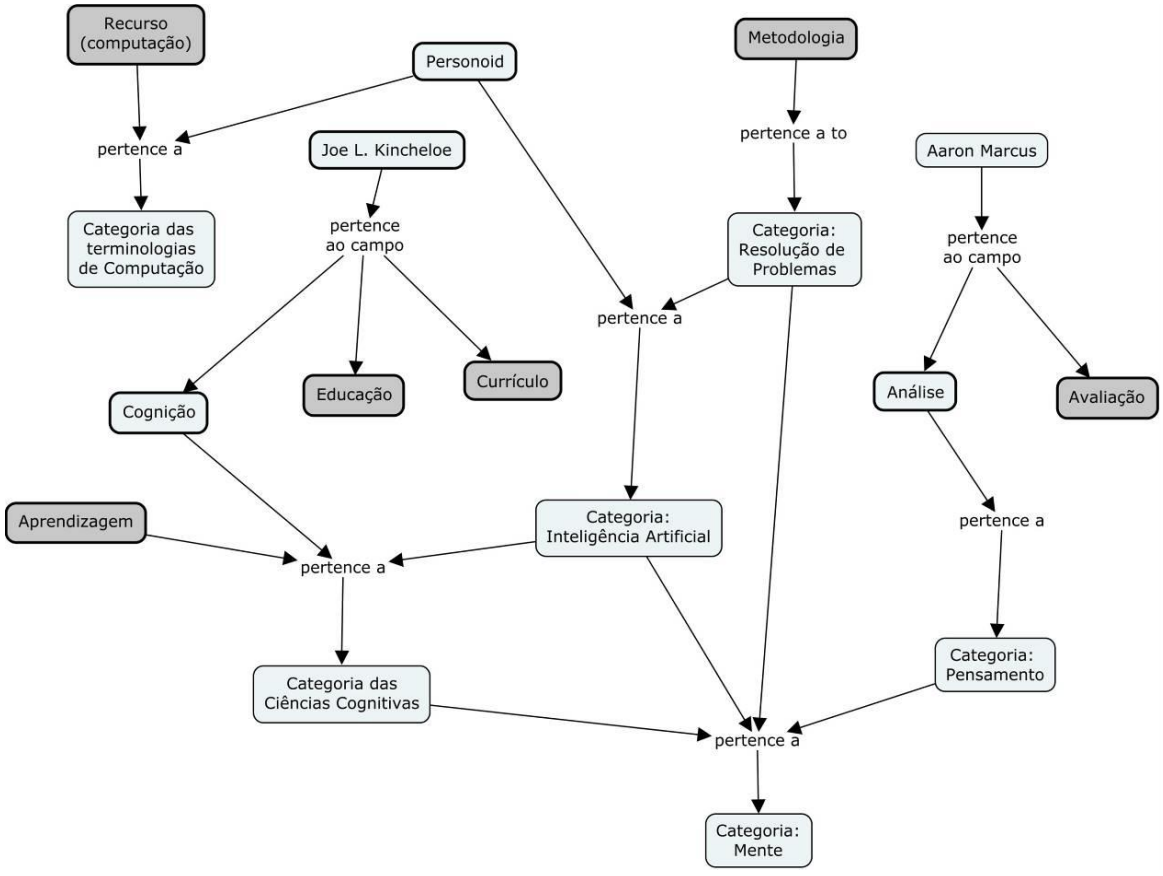
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 108 – Mapa resultante do usuário 5, a partir de três termos



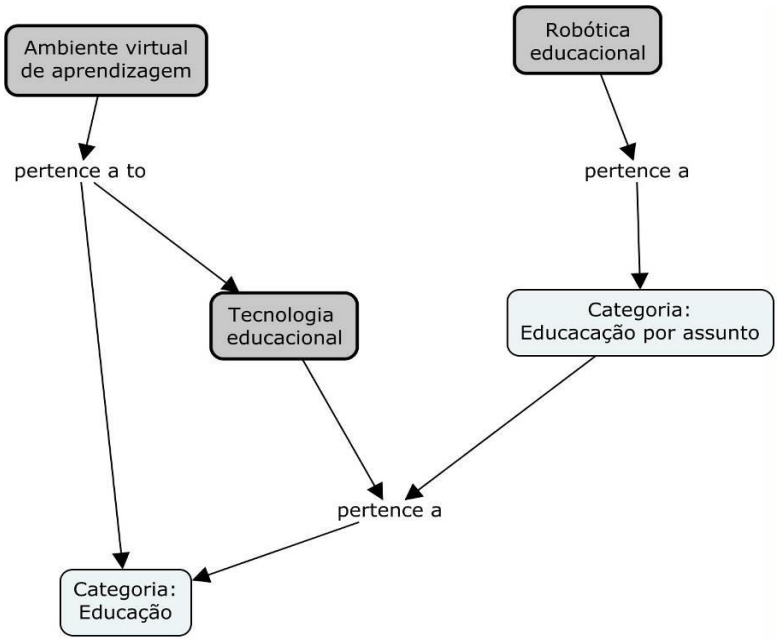
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 109 – Mapa resultante do usuário 5, a partir de seis termos



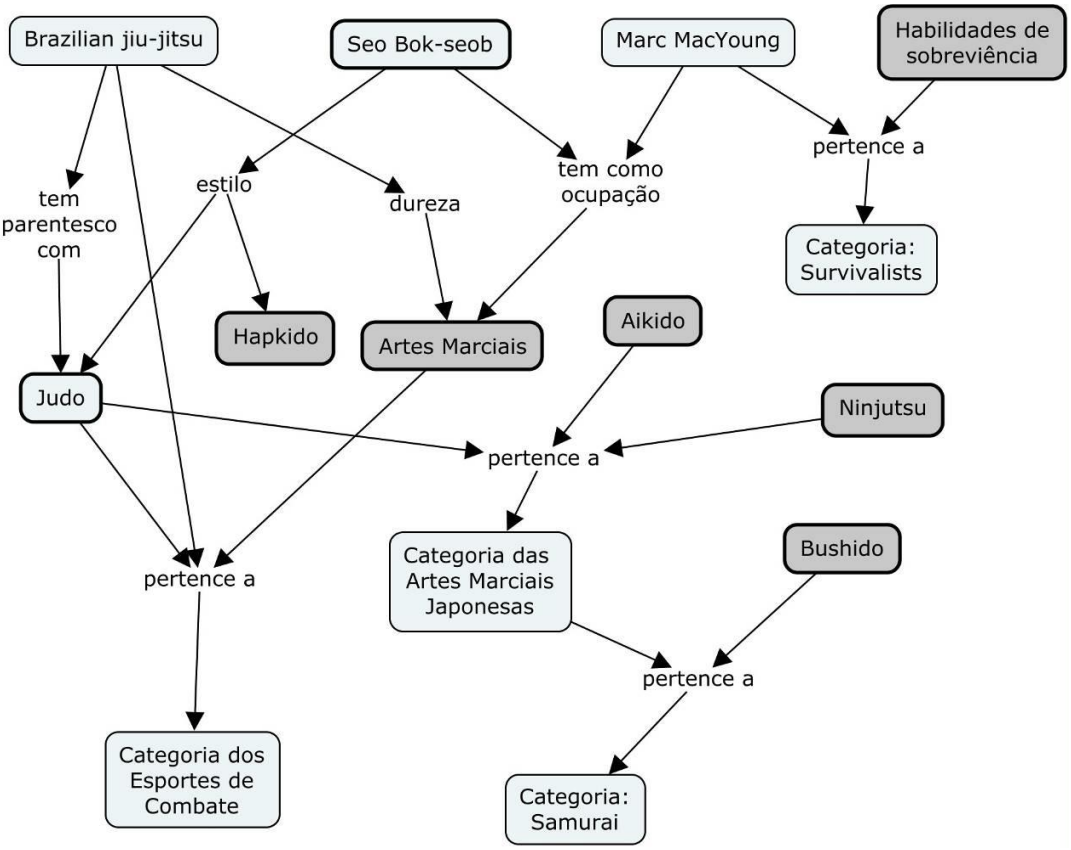
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 110 – Mapa resultante do usuário 6, a partir de três termos



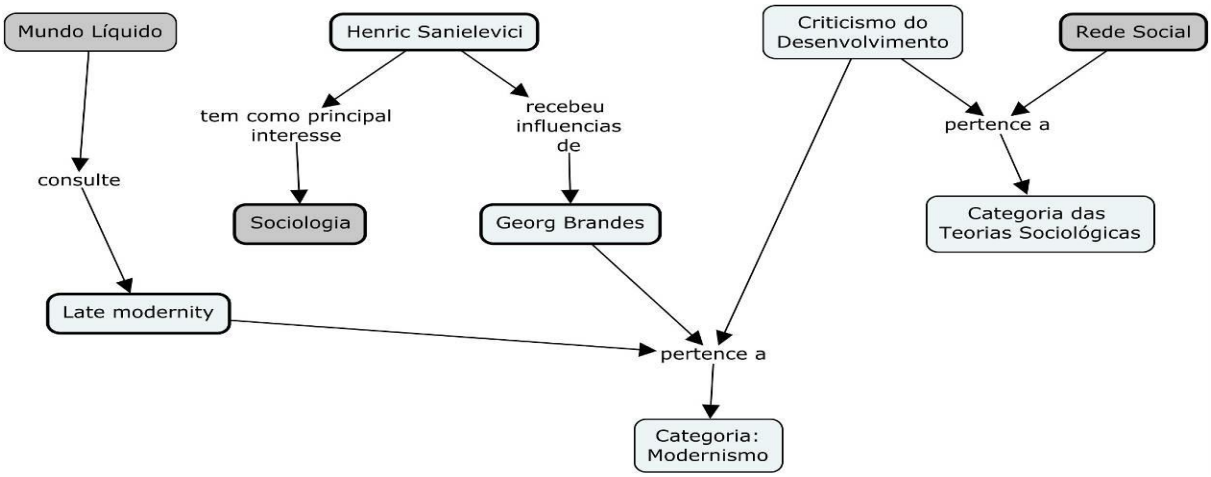
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 111 – Mapa resultante do usuário 6, a partir de seis termos



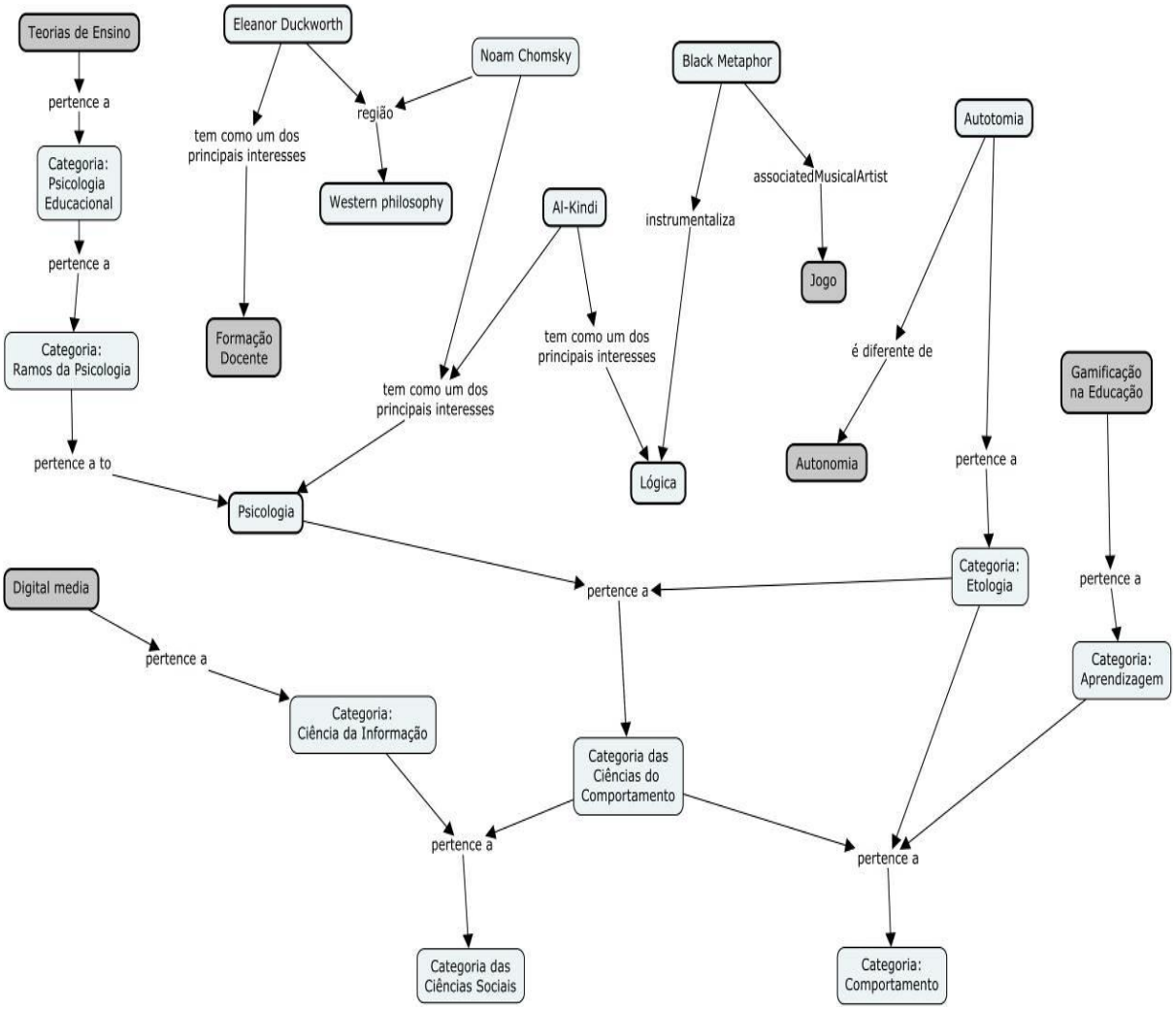
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 112 – Mapa resultante do usuário 7, a partir de três termos



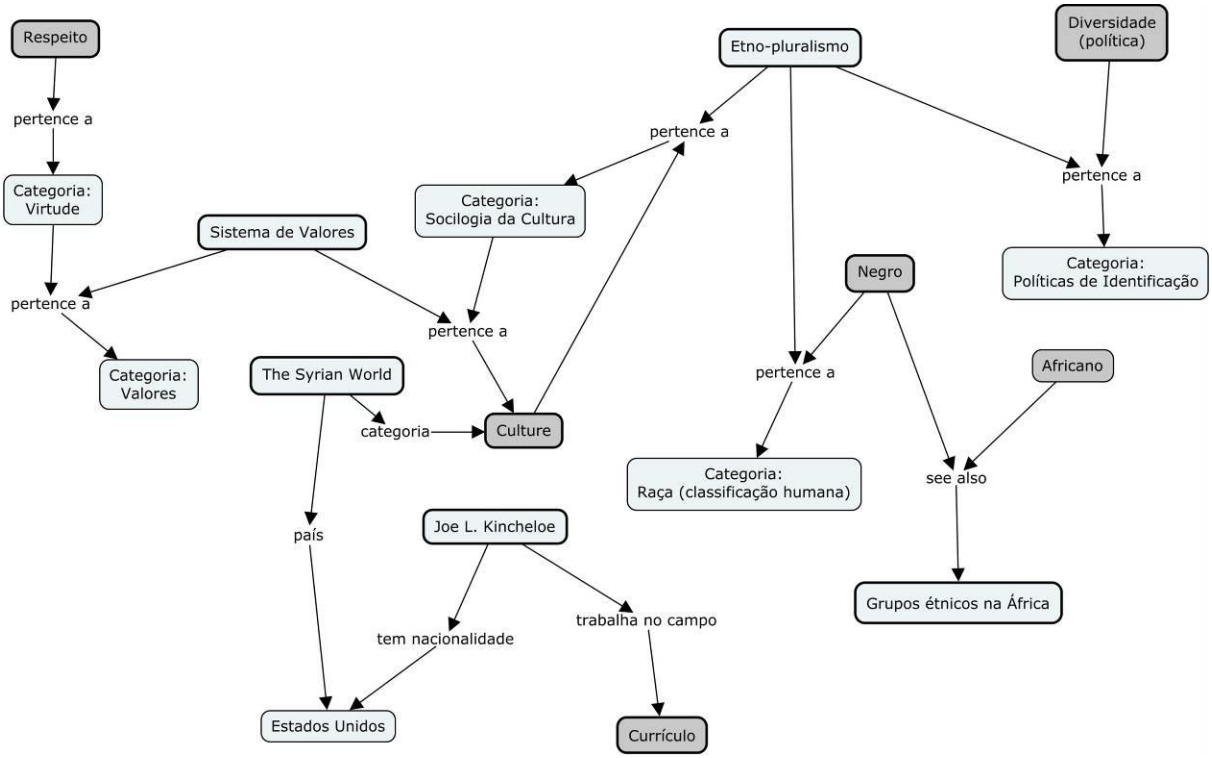
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 113 – Mapa resultante do usuário 7, a partir de seis termos



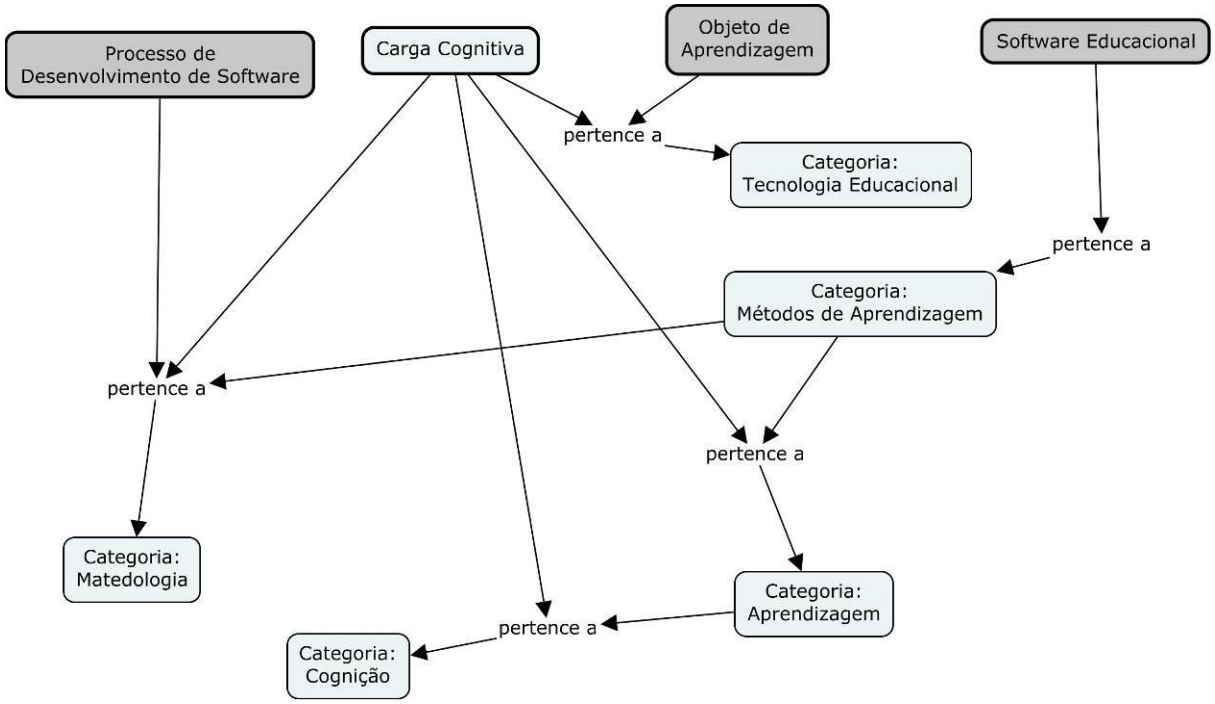
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 114 – Mapa resultante do usuário 8, a partir de seis termos



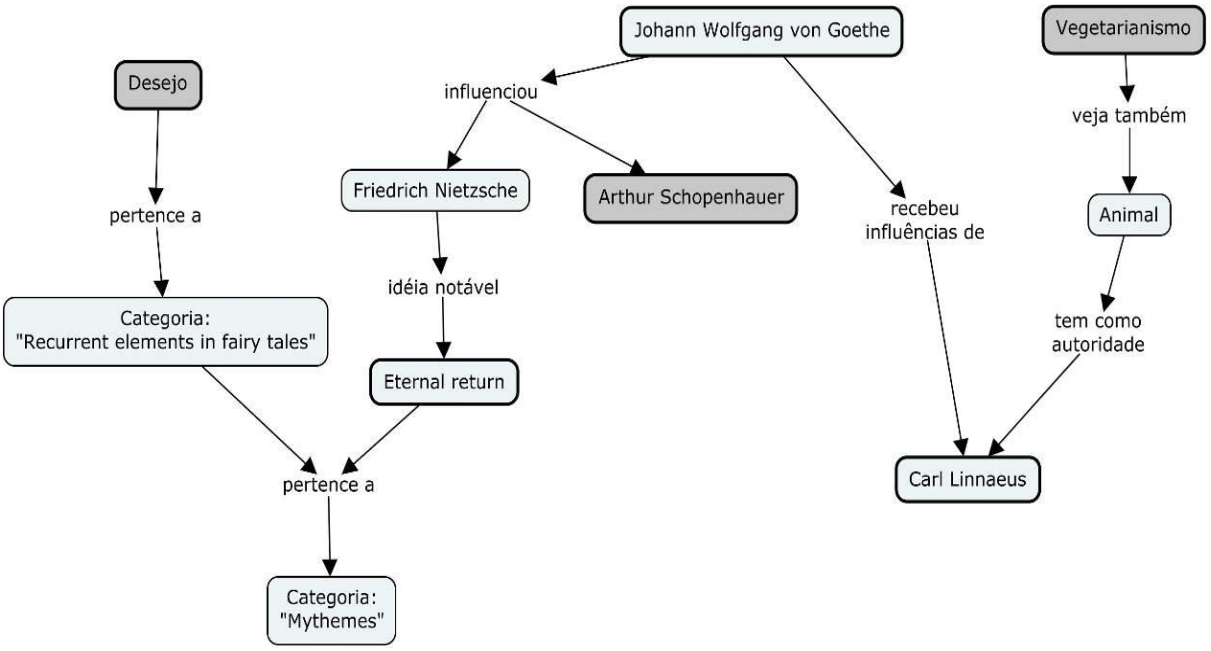
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 115 – Mapa resultante do usuário 9, a partir de três termos



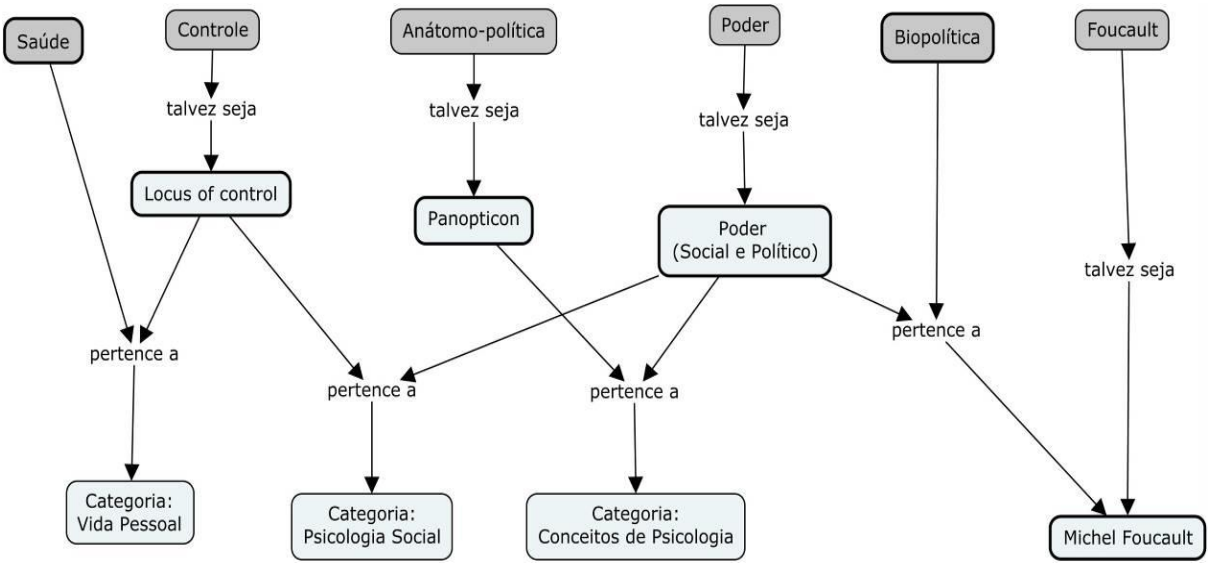
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 116 – Mapa resultante do usuário 10, a partir de três termos



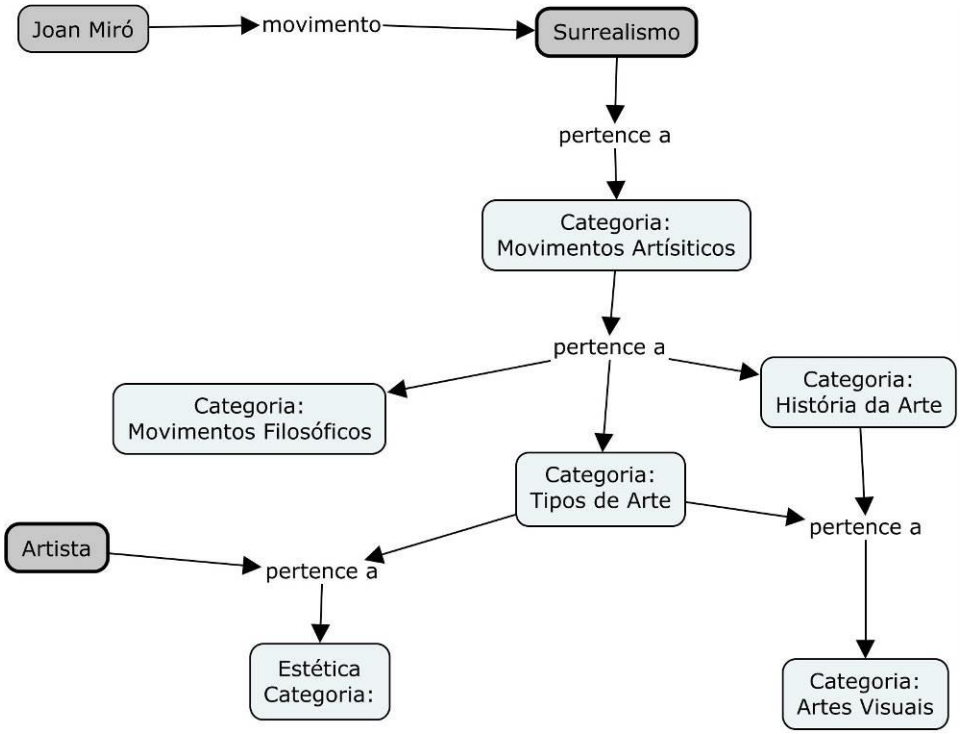
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 117 – Mapa resultante do usuário 10, a partir de seis termos



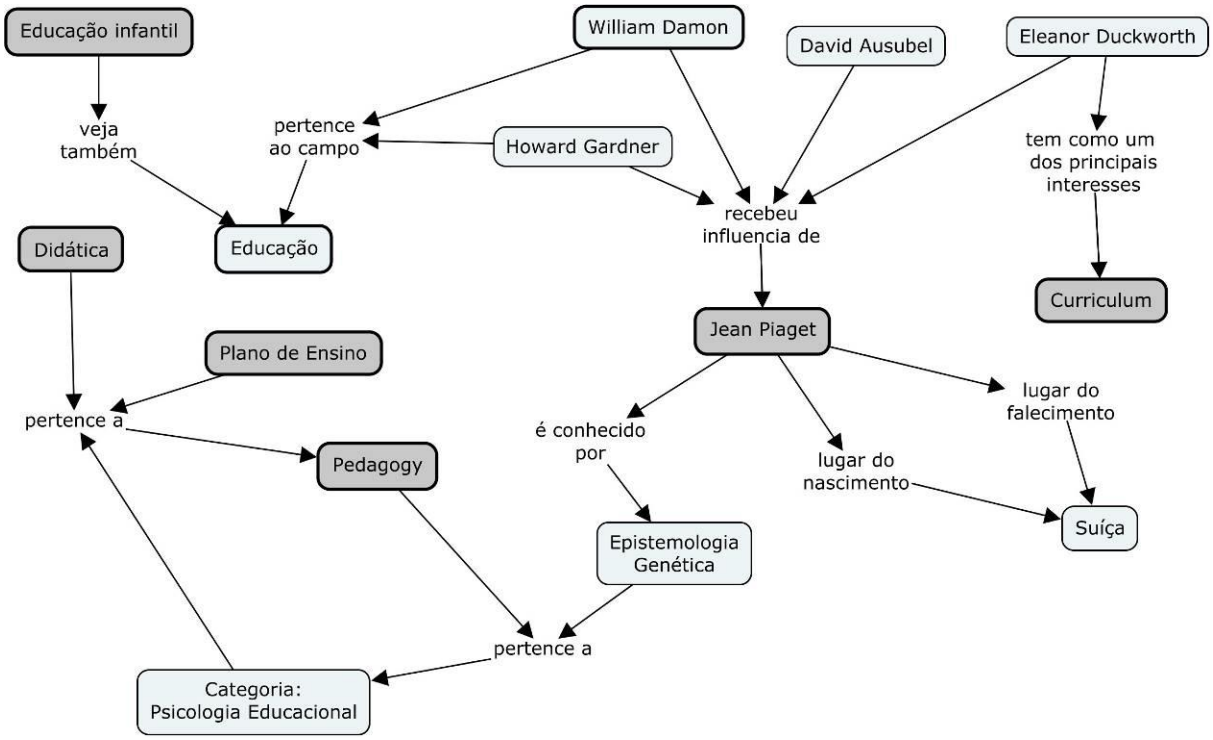
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 118 – Mapa resultante do usuário 12, a partir de três termos



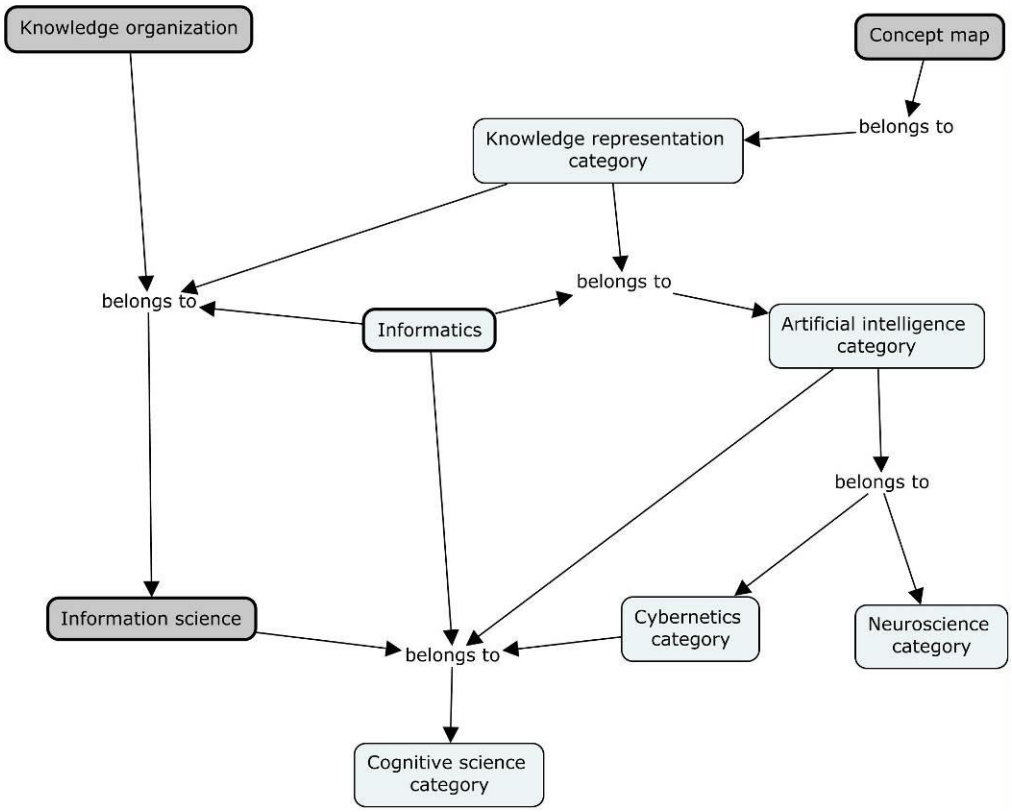
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 119 – Mapa resultante do usuário 12, a partir de seis termos



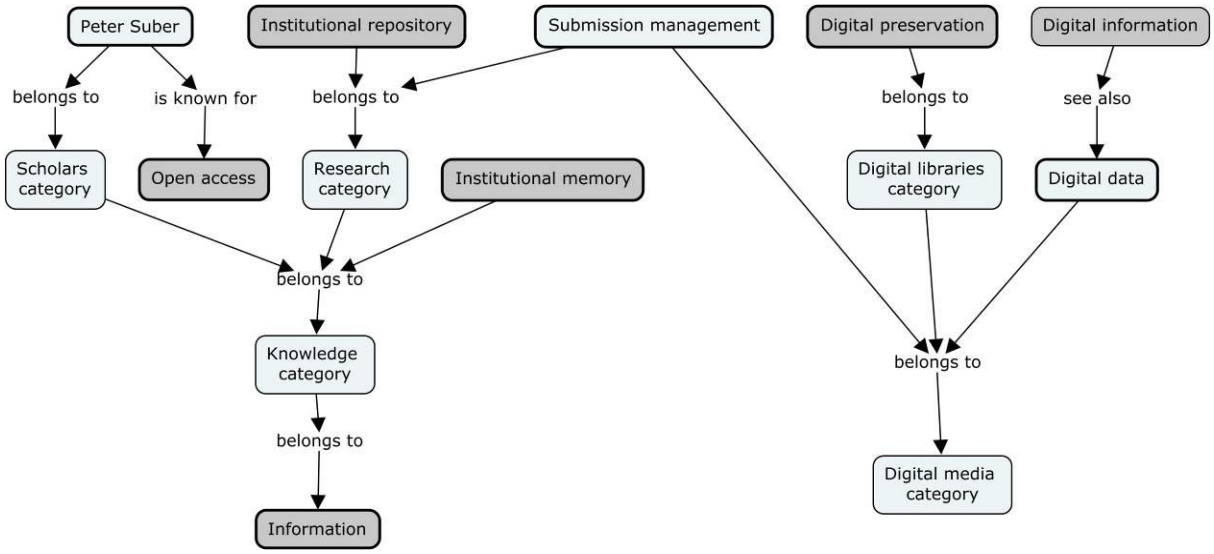
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 120 – Mapa resultante do usuário 13, a partir de três termos



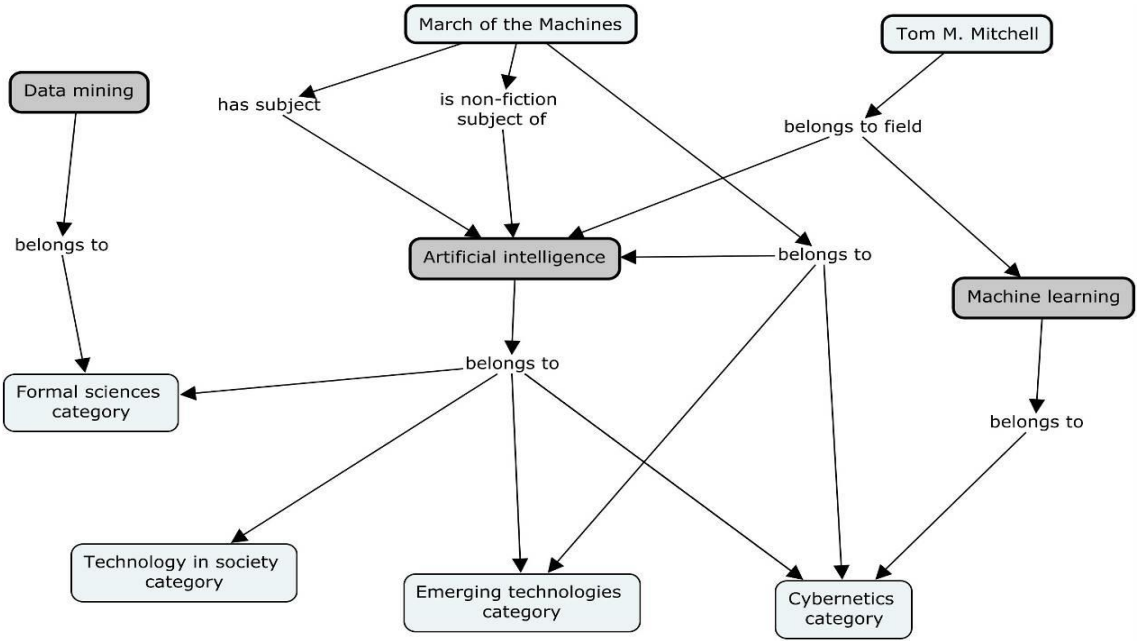
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 121 – Mapa resultante do usuário 13, a partir de seis termos



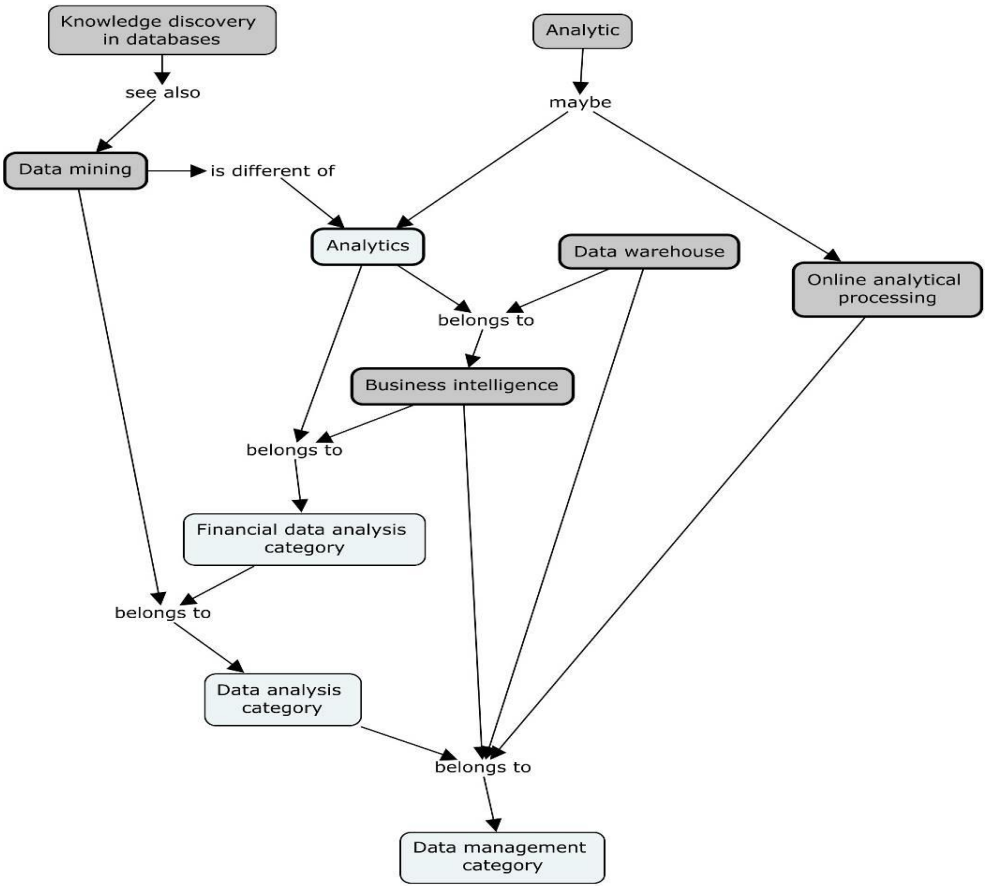
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 122 – Mapa resultante do usuário 14, a partir de três termos



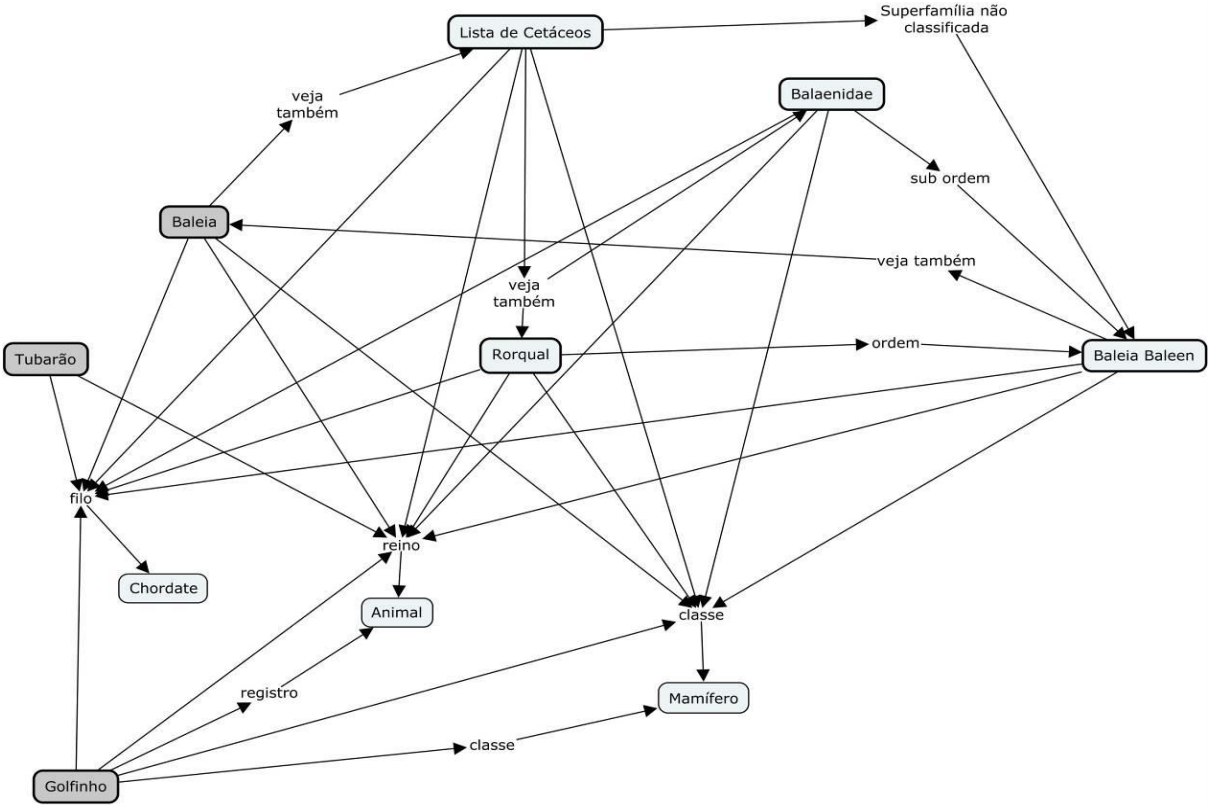
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 123 – Mapa resultante do usuário 14, a partir de seis termos



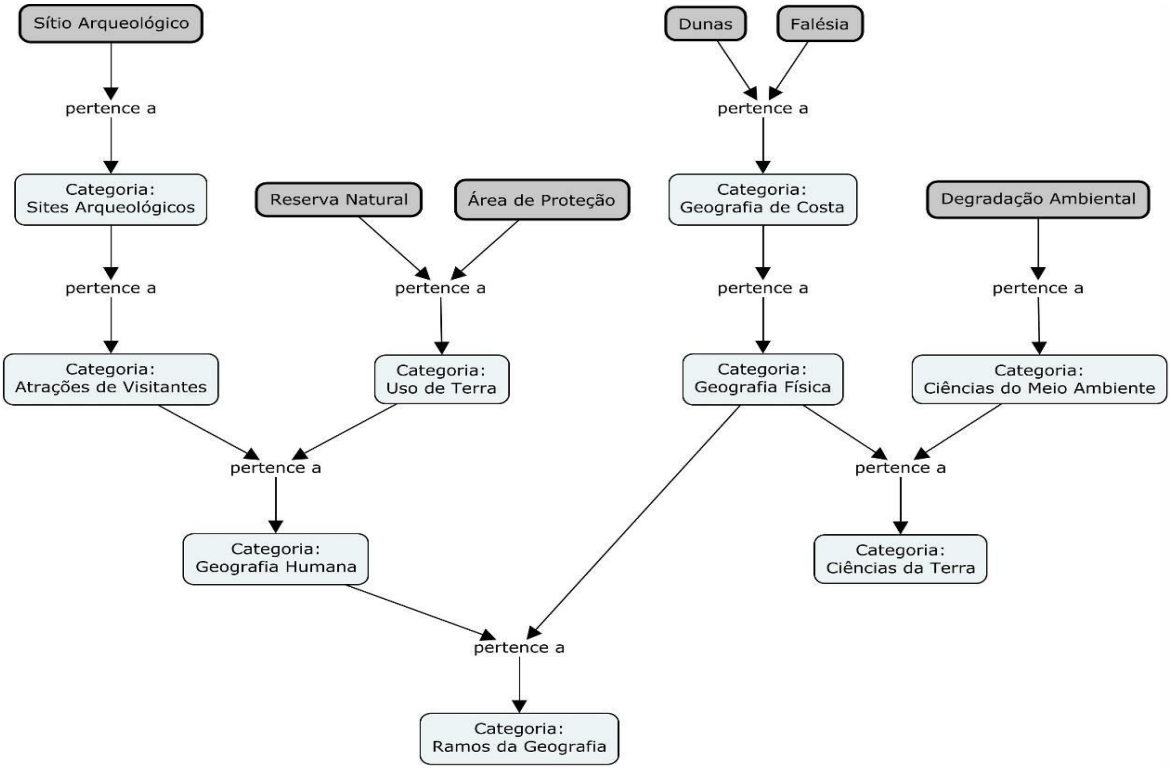
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 124 – Mapa resultante do usuário 15, a partir de três termos



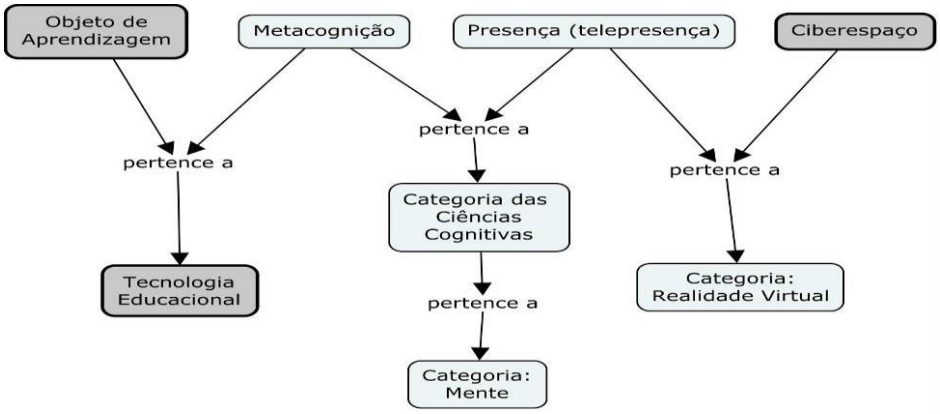
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 125 – Mapa resultante do usuário 15, a partir de seis termos



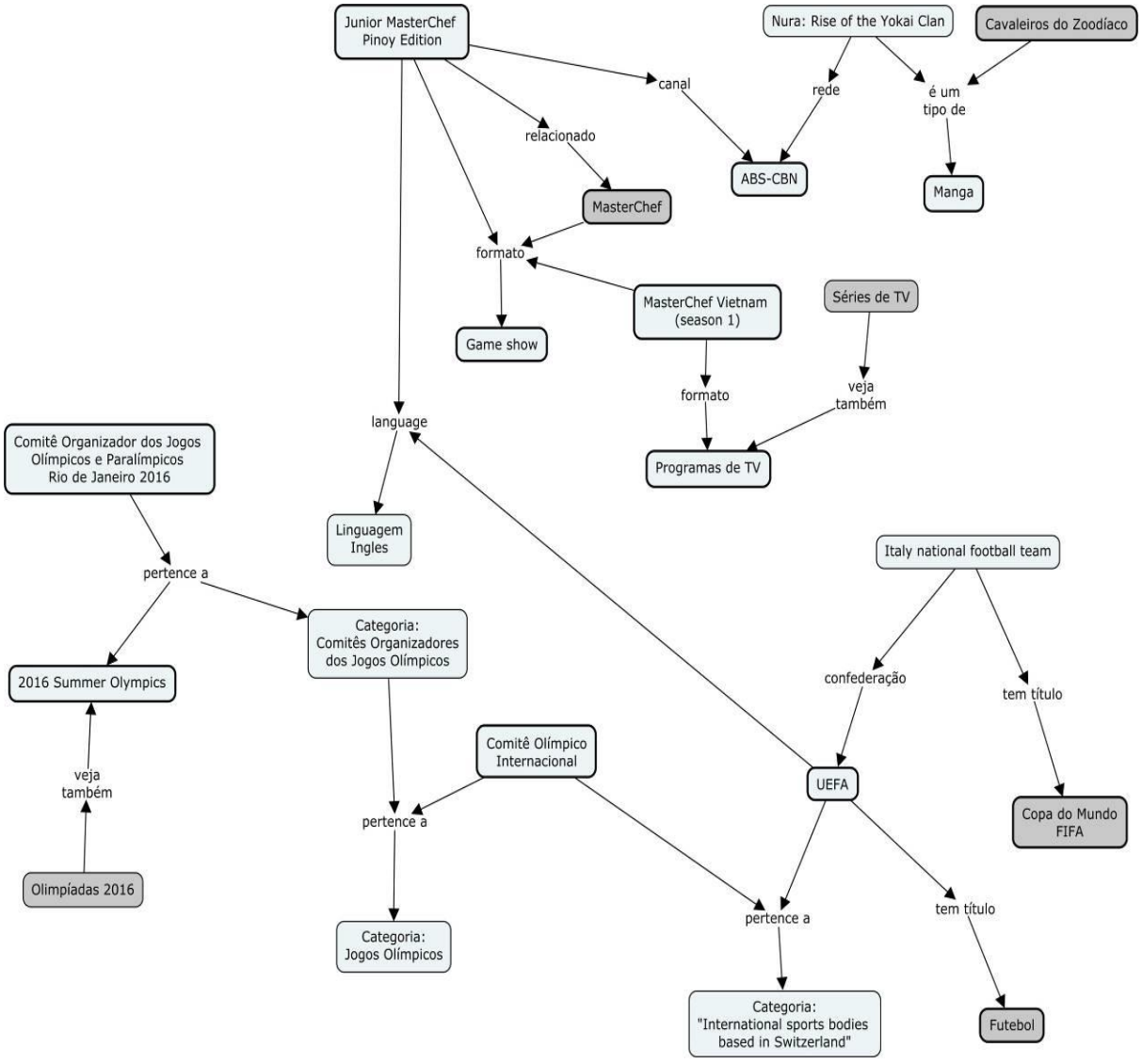
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 126 – Mapa resultante do usuário 16, a partir de três termos



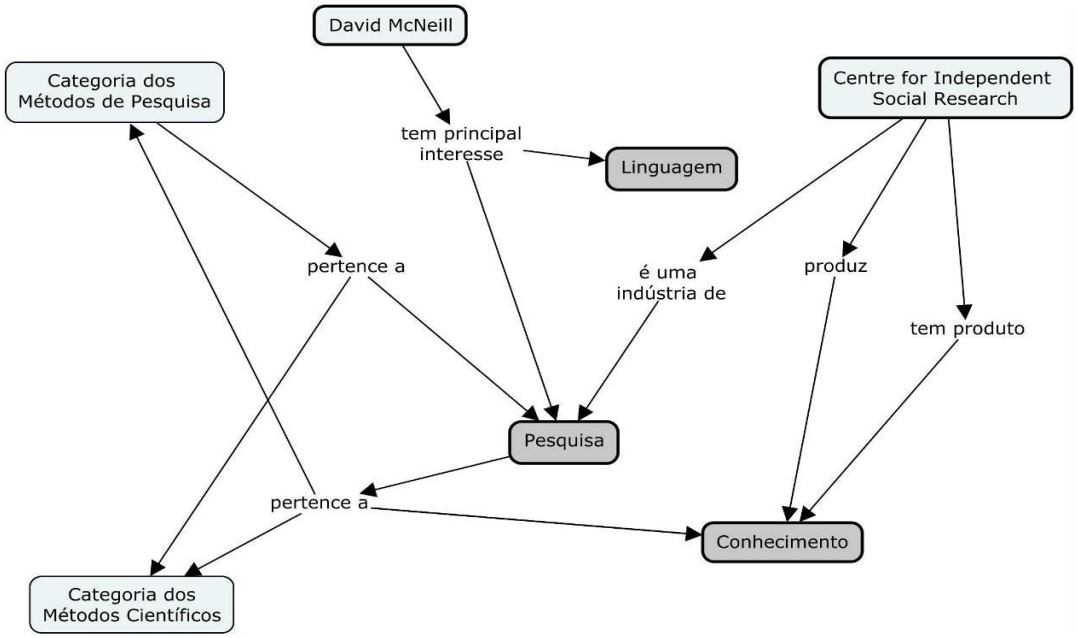
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 127 – Mapa resultante do usuário 16, a partir de seis termos



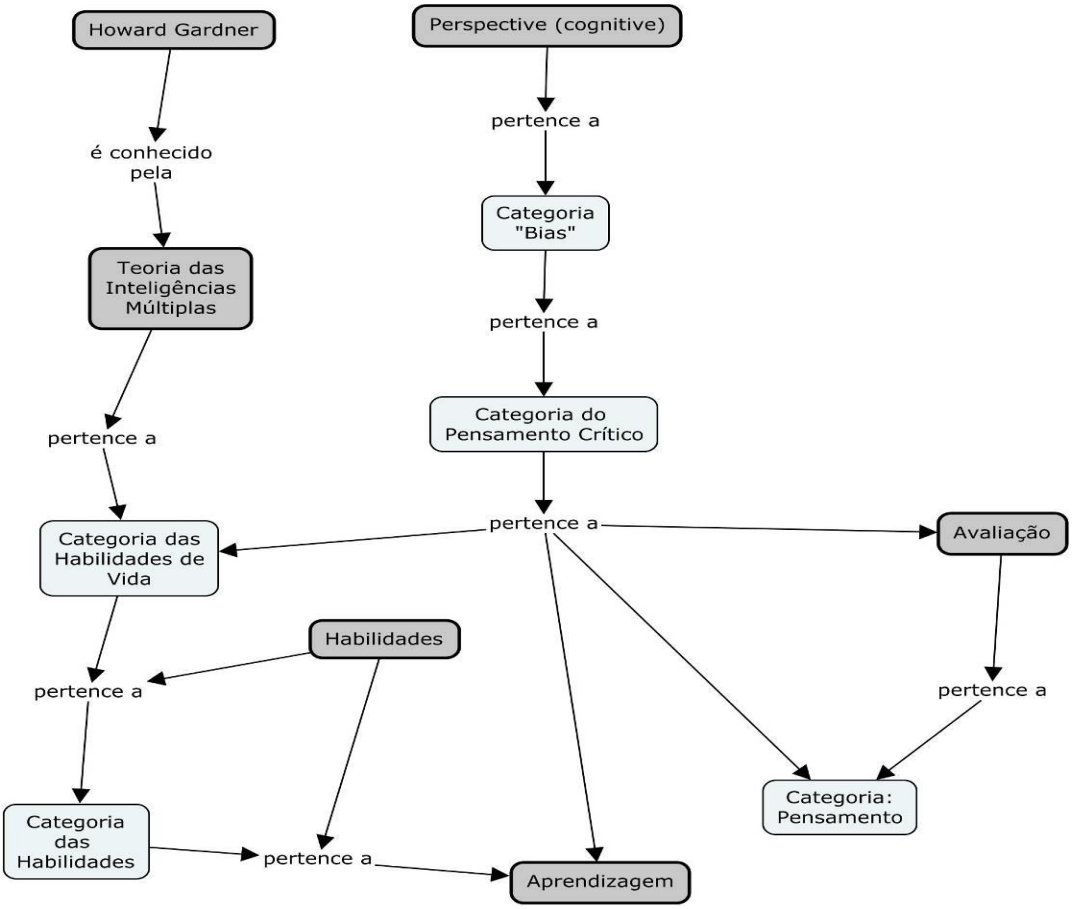
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 128 – Mapa resultante do usuário 17, a partir de três termos



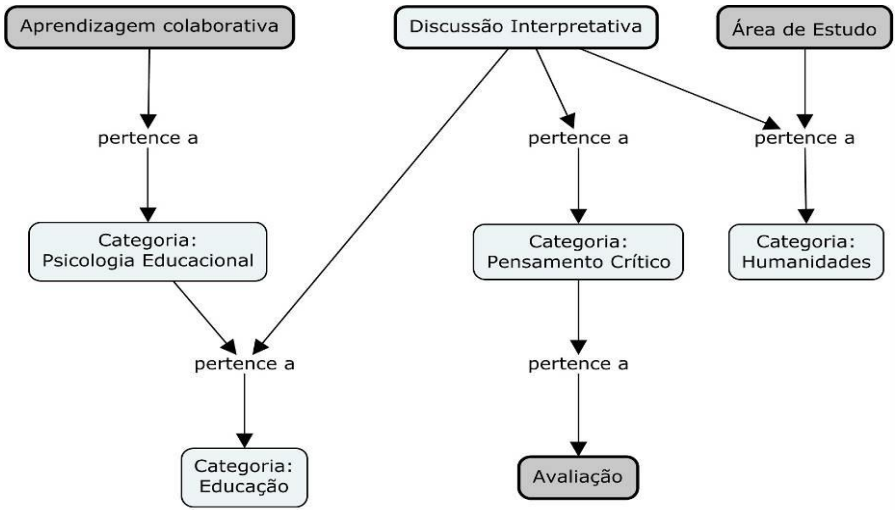
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 129 – Mapa resultante do usuário 17, a partir de seis termos



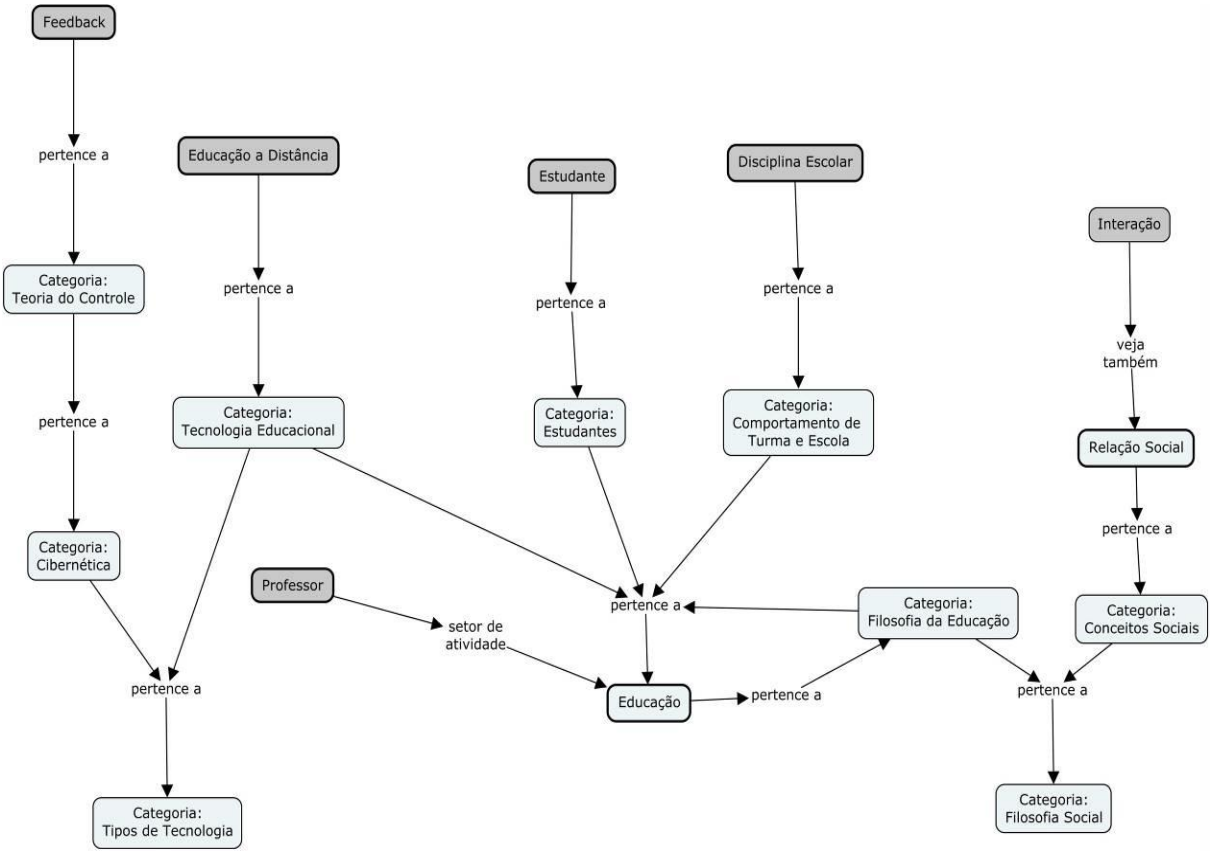
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 130 – Mapa resultante do usuário 18, a partir de três termos



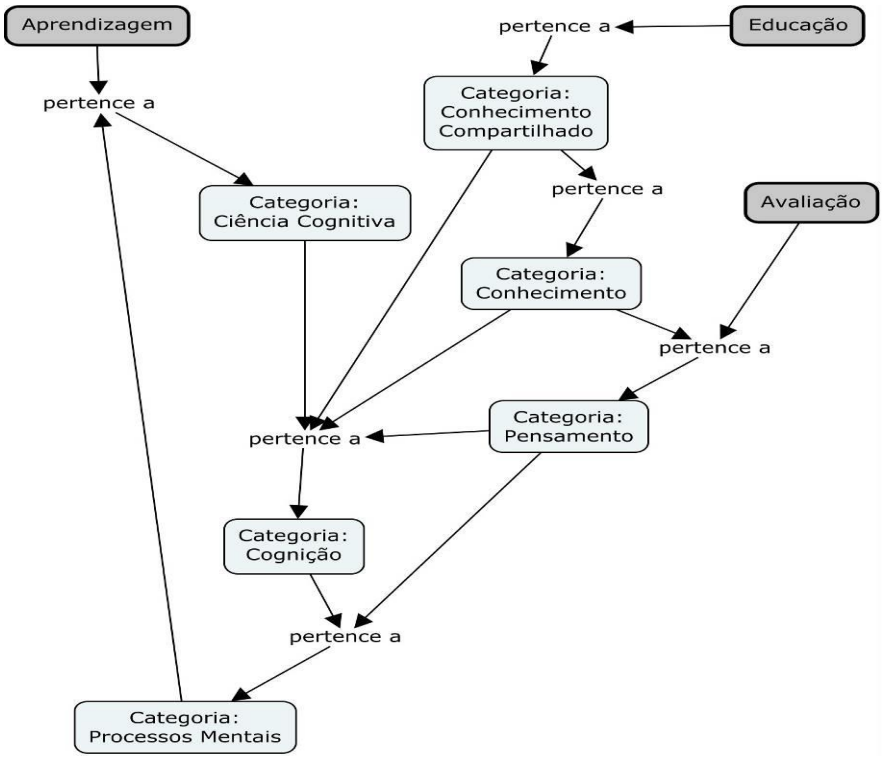
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 131 – Mapa resultante do usuário 18, a partir de seis termos



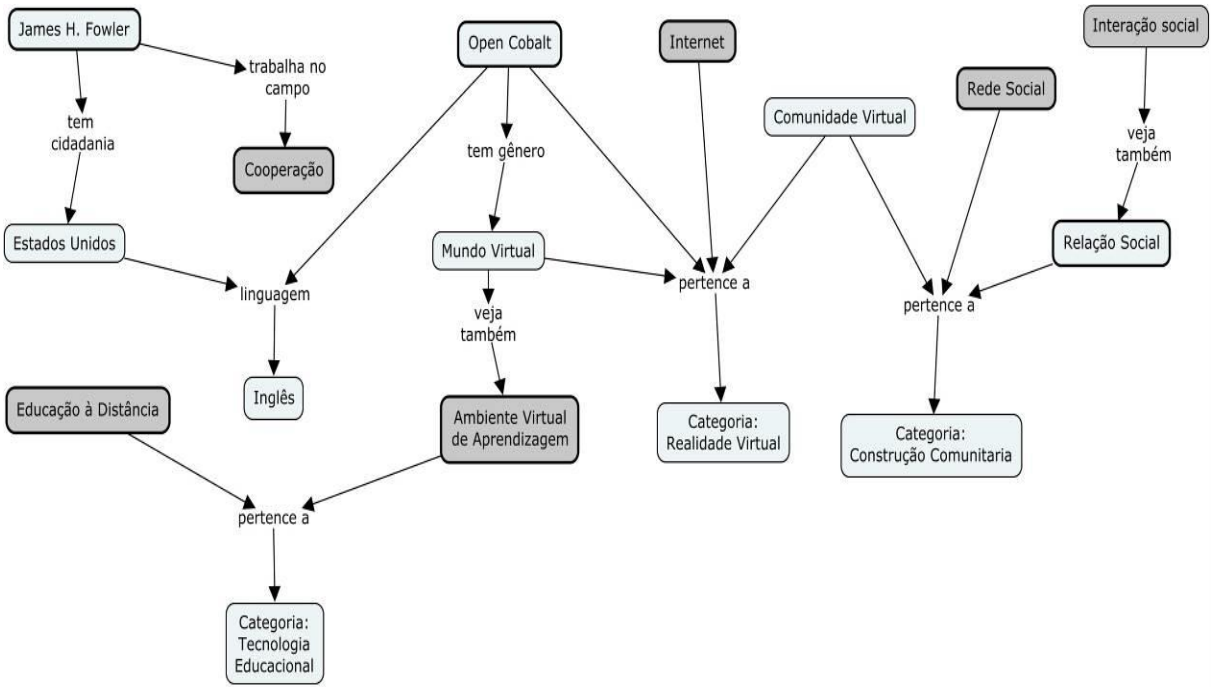
Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 132 – Mapa resultante do usuário 19, a partir de três termos



Fonte: Elaboração própria, por intermédio do protótipo desenvolvido

Figura 133 – Mapa resultante do usuário 19, a partir de seis termos



Fonte: Elaboração própria, por intermédio do protótipo desenvolvido