UNIVERSIDADE DE BRASÍLIA
Faculdade de Economia, Administração,
Contabilidade e Ciência da Informação e Documentação
Departamento de Ciência da Informação e Documentação

PRECISÃO NO PROCESSO DE BUSCA E RECUPERAÇÃO DA INFORMAÇÃO

ROGÉRIO HENRIQUE DE ARAÚJO JÚNIOR

Tese apresentada ao Departamento de Ciência da Informação e Documentação da Universidade de Brasília como requisito para obtenção do título de Doutor em Ciência da Informação

Professora Orientadora: Drª KIRA TARAPANOFF

Brasília, DF
2005

UNIVERSIDADE DE BRASÍLIA
Faculdade de Economia, Administração,
Contabilidade e Ciência da Informação e Documentação
Departamento de Ciência da Informação e Documentação

PRECISÃO NO PROCESSO DE BUSCA E RECUPERAÇÃO DA INFORMAÇÃO

ROGÉRIO HENRIQUE DE ARAÚJO JÚNIOR

Tese apresentada ao Departamento de Ciência da Informação e Documentação da Universidade de Brasília como requisito para obtenção do título de Doutor em Ciência da Informação

Professora Orientadora: Drª KIRA TARAPANOFF

Brasília, DF
2005

UNIVERSIDADE DE BRASÍLIA
Faculdade de Economia, Administração, Contabilidade e Ciência da Informação Documentação
Departamento de Ciência da Informação e Documentação
Programa de Pós-Graduação em Ciência da Informação

FOLHA DE APROVAÇÃO

Título: Precisão no Processo de Busca e Recuperação da Informação

Autor: Rogério Henrique de Araújo Júnior

Área de concentração: Transferência da Informação

Linha de pesquisa: Gestão da Informação e do Conhecimento

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

Tese aprovada em: 23/03/2005

Aprovado por:

Profª. Drª. Kira Maria Antonia Tarapanoff
Presidente - Orientador – (UnB/PPGCINF)

Prof. Dr. Jaime Robredo
Membro Interno – (UnB/PPGCINF)

Profª. Drª. Marisa Bräscher Basílio Medeiros
Membro Interno – (UnB/PPGCINF)

Prof. Dr. Miguel Filho Ferreira de Oliveira
Membro Externo (CEF)

Prof. Dr. Ulf Gregor Baranow
Membro Externo (UFP)

Prof. Dr. Emir Jose Suaiden
Suplente – (UnB/PPGCINF)

Para Patricia, Isabelle e Luiz Philippe com amor

Freqüentemente se diz que as experiências devem ser realizadas sem idéias preconcebidas. Isso não é possível; não somente seria tornar estéril toda a experiência, como também não o poderíamos fazer mesmo que o quiséssemos. Cada um carrega consigo sua concepção de mundo da qual não se pode desfazer assim tão facilmente. Somos obrigados a nos servir da linguagem, por exemplo, e nossa linguagem é toda modelada por idéias preconcebidas e não poderia ser diferente. E são idéias preconcebidas inconscientes, mil vezes mais perigosas do que as outras.

(**Jules-Henri Poincaré** em "A Ciência e a Hipótese", p. 116)

AGRADECIMENTOS

A elaboração deste trabalho só foi possível devido a colaboração de inúmeras pessoas. A todas apresento os meus agradecimentos. Agradeço ainda, e em especial:

Ao Programa de Pós-graduação em Ciência da Informação e Documentação na pessoa da Professora Sely Maria de Souza Costa, pelo apoio, incentivo e amizade.

À Professora Eliane Braga de Oliveira, que me substituiu na Coordenação do Curso de Arquivologia na fase mais importante de elaboração do trabalho.

Aos profissionais da Caixa que entusiasticamente apoiaram-me em tudo que foi necessário: Rosane Helena C. P. de Araújo, Raimundo Nonato de Sousa, Dionne Benjamim e a todos os colaboradores da Cedin, os meus sinceros agradecimentos.

À Aloísio M. Carvalho por viabilizar a parceria entre a Caixa, UnB e a Policentro disponibilizando e adaptando o *software* BR/*Search* para a montagem do protótipo usado no trabalho.

À Carlos Henrique Ferreira de Araújo pelas sugestões na definição da amostra selecionada para o estudo.

Ao Professor Emir José Suaiden pela confiança e amizade em mim depositados.

Ao Professor Miguel Filho Ferreira de Oliveira que além de aceitar o convite para participar da banca examinadora, contribuiu com suas observações para o aprimoramento do trabalho.

Ao Professor Jaime Robredo pelo exemplo de dedicação a uma área de pesquisa, que pude comprovar, espinhosa, mas gratificante. Agradeço pelas sugestões, incentivo e participação na banca examinadora.

À Professora Marisa Bräscher Basílio Medeiros pela confiança depositada no meu trabalho, pelo incentivo e por emprestar o seu tempo para orientar-me na formulação da metodologia do trabalho.

Ao Professor Ulf Gregor Baranow pela paciência e grande generosidade ao apontar caminhos viáveis para a construção de toda a pesquisa. Agradeço a orientação e a participação na banca de examinadora, consciente do esforço empreendido para tanto.

À Professora Kira Tarapanoff, que desde o primeiro momento aceitou orientar-me mais uma vez. Sem os seus ensinamentos, sua experiência, sua inteligência, profissionalismo e dedicação este empreendimento científico teria sido inviável.

À minha cheia de luz e amada Patricia Marie Jeanne Cormier, pela revisão dos originais e sugestões de aperfeiçoamento do texto.

RESUMO

ARAÚJO JR., R. H. de. *Precisão no processo de busca e recuperação da informação*. Brasília : Universidade de Brasília, 2005 (tese de doutorado em Ciência da Informação).

Esta pesquisa trata da comparação entre a indexação manual e a ferramenta de mineração de textos, por meio da análise do índice de precisão de resposta no processo de busca e recuperação da informação. O estudo de caso escolhido para o desenvolvimento da pesquisa foi o Centro de Referência e Informação em Habitação - Infohab, cuja base de dados sobre habitação, saneamento e urbanização foi indexada de forma manual por bibliotecários da Caixa Econômica Federal - CAIXA, com base em uma lista de palavras-chave. Houve o desenvolvimento de um protótipo cujos itens bibliográficos correspondem às teses e dissertações contidas no Infohab, o que permitiu a aplicação do *software* BR/*Search* para a execução da mineração de textos. As pesquisas realizadas no Infohab e no protótipo foram feitas a partir da demanda de especialistas da CAIXA nos assuntos contidos na base. Conclui que não há ganhos significativos na precisão ao se aplicar a ferramenta de mineração de textos em relação à indexação manual. Não obstante, a utilização do índice de precisão deverá sempre propiciar parâmetros para a melhoria contínua da resposta obtida dos sistemas de recuperação da informação, cuja avaliação de performance caberá ao usuário, que vai definir, em nome de sua necessidade de informação, o que é útil ou inútil dentre toda a informação recuperada.

PROCESSO DE RECUPERAÇÃO DA INFORMAÇÃO
ÍNDICE DE PRECISÃO
REVOCAÇÃO
PROCESSO DE INDEXAÇÃO
MINERAÇÃO DE TEXTOS

ABSTRACT

ARAÚJO JR., R. H. de. *Precisão no processo de busca e recuperação da informação*. Brasília : Universidade de Brasília, 2005 (tese de doutorado em Ciência da Informação)

This research deals with the comparison between manual indexing and the text mining tool, using the analysis of reply precision rate in the information retrieval process. The case study selected for this research was the Centro de Referência e Informação em Habitação – Infohab. The center database on habitation, sanitation and urbanization was manually indexed by the librarians of Caixa Econômica Federal – CAIXA, using a list of key words. A prototype was developed, containing bibliographic references that corresponded to the theses and dissertations of Infohab, which allowed text mining application by BR/Search software. The researches performed on the prototype and in Infohab were demanded by specialists of CAIXA in database subjects. The research evidenced that there are no significant profits in the precision rate in the applications of text mining tool in relation to the manual indexing. However, the use of precision rate will always provide parameters to the improvement of information retrieval, to be assessed by the user who should define, according to the information need, what is useful or useless amongst the information retrieved.

INFORMATION RETRIEVAL PROCESS
PRECISION RATE
RECALL
INDEXING PROCESS
TEXT MINING

SUMÁRIO

RELAÇÃO DE ABREVIATURAS E SIGLAS

ANTAC - Associação Nacional de Tecnologia do Ambiente Construído

CEDIN – Centralizadora de Informação e Documentação

CAIXA - Caixa Econômica Federal

CRM – *Customer Relationship Management*

DIPUP – Diretoria de Parcerias e Apoio ao Desenvolvimento Urbano

DSI – Disseminação Seletiva da Informação

FCS – Fatores Críticos de Sucesso

FINEP - Financiadora de Estudos e Projetos

GED - Gerenciamento Eletrônico de Documentos

GEMAC - Gerência Nacional de *Marketing* Interno

GEPAD – Gerência Nacional de Normas e Padrões de Engenharia e Trabalho Social

GERED – Gerência Nacional de Gestão de Rede de Filiais

GEURB - Gerência Nacional de Prestação de Serviços em Desenvolvimento Urbano

IBICT – Instituto Brasileiro de Informação em Ciência e Tecnologia

Infohab - Centro de Referência e Informação em Habitação

ISP - *Information Search Process*

IR- *Information Retrieval*

KDD - *Knowledge Discovery in Databases*

KDT – *Knowledge Discovery in Text*

MCT - Ministério da Ciência e Tecnologia

PBRI – Processo de Busca e Recuperação da Informação

PPGC – Programa de Pós-Graduação em Computação da UFRGS

SRI – Sistema de Recuperação da Informação

UFBA – Universidade Federal da Bahia

UFF - Universidade Federal Fluminense

UFMG – Universidade Federal de Minas Gerais

UFRGS – Universidade Federal do Rio Grande do Sul

UFRJ – Universidade Federal do Rio de Janeiro

UFSC – Universidade Federal de Santa Catarina

UFSCAR – Universidade Federal de São Carlos

UnB - Universidade de Brasília

USP – Universidade de São Paulo

VIURB – Vice-Presidência de Desenvolvimento Urbano e Governo

RELAÇÃO DE TABELAS

RELAÇÃO DE GRÁFICOS

RELAÇÃO DE QUADROS

RELAÇÃO DE FIGURAS

1- INTRODUÇÃO

Qualquer organização hoje tem seus métodos e processos, bem como a sua ação administrativa, amplamente apoiados na gestão do seu fluxo informacional. Junta-se a esta questão, uma série de elementos, desde a gestão das tarefas diárias que fazem parte de planos de ação, até situações globais relacionadas às estratégias organizacionais. Em todas estas etapas, o gerenciamento do fluxo da informação tem se constituído elemento indispensável, não apenas para o sucesso do empreendimento, mas também para a compreensão clara e a comunicação entre todas as áreas, ou subsistema da organização.

Neste sentido, as corporações, além do investimento tradicional no controle dos sistemas de informação focando seus elementos de entrada, processamento e saída, têm buscado soluções para a recuperação da informação. Nesta tarefa, cada vez mais considerada como fundamental para o processo decisório, a utilização da inteligência artificial, a partir de aplicações comerciais, tem contribuído para a sua otimização.

Segundo Wiston (1997), as possibilidades de aplicação da inteligência artificial na recuperação da informação estão focadas nos sistemas de Intranet e Internet que filtram gigantescas quantidades de informações e apresentam resultados em formatos legíveis e simples; na tecnologia que utiliza linguagem natural para recuperar qualquer tipo de informação em linha, quer seja texto ou imagem; e nas ferramentas de mineração de dados e mineração de textos.

A recuperação da informação deve ser entendida, segundo Belkin & Croft (1987), como o processo de localizar documentos e itens de informação que tenham sido objeto de armazenamento, para permitir o acesso dos usuários aos itens de informação, objetos de uma solicitação, ou seja, a recuperação da informação se dá pela comparação do que se solicitou com o que está

armazenado. Este processo possui como elementos vitais a indexação e o armazenamento.

O armazenamento dos itens de informação em uma base de dados só pode ser feito se os documentos que vão compor o sistema passarem por uma análise meticulosa de seus conteúdos. Neste momento, a indexação entra em cena para viabilizar a escolha dos termos que irão representar os conteúdos dos documentos, que por sua vez, serão imprescindíveis na recuperação da informação. Este princípio é universal na gestão dos sistemas de informação. Para Rowley (2002), todos os sistemas de recuperação da informação podem ser compreendidos como se fossem formados por três etapas: indexação, armazenamento e recuperação.

Partindo da premissa que a análise do documento é uma significativa contribuição para a comunicação e o fluxo da informação em qualquer organização e para qualquer sistema de recuperação da informação, o tema da pesquisa proposto trata da comparação entre indexação manual e ferramenta de mineração de textos, por meio da análise do índice de precisão de resposta no processo de busca e recuperação da informação. O estudo de caso escolhido para o desenvolvimento da pesquisa foi o Centro de Referência e Informação em Habitação – Infohab.

O Infohab é liderado pela Associação Nacional de Tecnologia do Ambiente Construído – ANTAC e tem por finalidade a captação, seleção e divulgação de toda a informação a cerca da tecnologia do ambiente construído, sobretudo sobre a habitação de interesse social, englobando a sua produção, manutenção e uso.

O ambiente construído envolve todas as atividades, recursos, conhecimento, expertises, experiências, tecnologia, equipamentos, instrumentos, mão-de-obra e mercado relacionados à habitação e às políticas públicas sobre

saneamento, urbanização, além das questões técnicas que envolvem o estatuto das cidades.

O foco do estudo de caso incide sobre uma das competências do Infohab, a manutenção de uma base de dados atualizada com referências dos resultados de pesquisas, legislação federal, estadual/municipal, normas pertinentes, levantamentos governamentais e demais tipologias de documentação.

O universo da pesquisa é representado por 1520 documentos da Caixa Econômica Federal (CAIXA) inseridos na base de dados do Infohab e, como amostra, o acervo de teses e dissertações inseridas no Infohab pela Centralizadora de Documentação e Informação (CEDIN), vinculada à Gerência Nacional de *Marketing* Interno da CAIXA (GEMAC).

1.1- Tema e problema

1.1.1- Tema

Estudo comparado entre a mineração de textos e a indexação manual, por meio da análise do índice de precisão de resposta no processo de busca e recuperação da informação: o caso do Centro de Referência e Informação em Habitação – Infohab.

1.1.2- Problema

O trabalho proposto consiste em avaliar se no processo de busca e recuperação da informação a mineração de textos traz ganho no índice de precisão em relação à lista de palavras-chave utilizadas na indexação manual por bibliotecários da Caixa Econômica Federal – CAIXA.

Além disso, interessa ao escopo da investigação, verificar a viabilidade de propor uma sistemática de uso dos termos gerados a partir de mineração de textos para auxiliar o processo de indexação manual, no aumento do índice de precisão de resposta na recuperação da informação.

O estudo trata da comparação entre a mineração de textos e a indexação manual, por meio da análise do índice de precisão de resposta no processo de busca e recuperação da informação. O estudo de caso incidiu sobre o Centro de Referência e Informação em Habitação – Infohab, cuja base de dados sobre habitação, saneamento e urbanização foi indexada manualmente por bibliotecários da CAIXA, com base em uma lista de palavras-chave. Houve o desenvolvimento de um protótipo cujos itens bibliográficos correspondem às teses e dissertações contidas no Infohab, o que possibilitou o emprego do *software* BR/*Search* para a execução da mineração de textos.

Para melhor compreensão da problema, cabe definir mineração de textos. A mineração de textos consiste na extração de informações sobre tendências ou padrões em grandes volumes de documentos textuais, onde uma amostra significativa de informações é avaliada em textos contidos em bases textuais e em fontes de informação em linha (Polanco & François, 2000).

As bases textuais são coleções de documentos em linguagem natural, sem formato pré-definido para seus conteúdos, como acontece com as bases de dados. Dividem-se em:

I. Bases textuais estruturadas cujo conteúdo é estruturado de acordo com a sua localização no documento. Como exemplos têm-se: relatórios policiais, relatórios de instituições financeiras, ou seja, o conteúdo pode variar, mas a estrutura do documento é pré-definida; e

II. Informações não estruturadas, onde têm-se como exemplo os relatórios, publicações e a maioria dos documentos textuais (Trybula, 1999).

Na Figura 1, a seguir, são apresentados os fatores de influência nos resultados de busca e recuperação da informação em uma base de dados:

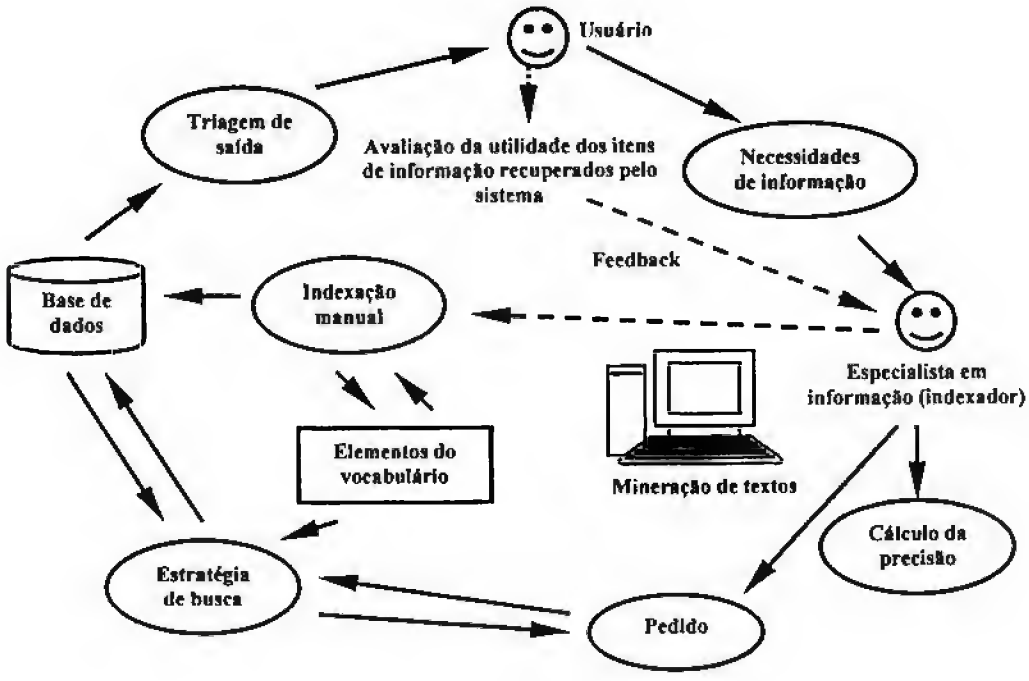


Figura 1 - Fatores de influência nos resultados de busca em uma base de dados (Fonte: adaptado de LANCASTER, 1998)

Neste modelo adaptado de Lancaster (1998), podemos visualizar as peças básicas que compõem a problemática apresentada, onde o usuário, a partir de suas necessidades de informação, dá início ao processo de busca e recuperação da informação.

O especialista em informação, ao qual cabe a responsabilidade de indexador, se vale da compreensão da demanda para trazer da base de dados, na forma do pedido, a resposta o mais adequada possível às necessidades do usuário. Além disto, Lancaster (1998) ressalta que a qualidade da estratégia de busca e o vocabulário são fatores importantes para a atividade.

Todavia, temos também as questões da precisão e da qualidade da própria base de dados, sem contar que o indexador (especialista em informação) depende dos termos autorizados no vocabulário, a fim de lograr êxito na indexação que deve alimentar e impactar a base de dados.

1.2- Propósito da pesquisa

O propósito do estudo é comparar o resultado da aplicação da mineração de textos com o resultado da aplicação da recuperação de itens (recuperação informacional de textos manualmente indexados), em função do índice de precisão de resposta no processo de busca e recuperação da informação.

Desta forma, pretende-se com este estudo:

I. Verificar se na indexação manual, o índice de precisão resultante do processo de busca e recuperação da informação é superado com o uso de ferramenta de mineração de textos;

II. Verificar se a ferramenta de mineração de textos pode ser convertida em ferramenta de indexação a partir da extração automática de termos, com o auxílio do julgamento dos indexadores na seleção de termos a serem utilizados na representação do conteúdo dos documentos em futuras pesquisas; e

III. Se a mineração de textos poderá apoiar a construção e/ou manutenção do tesauro, que é gerado e usado na indexação manual.

Com o trabalho pretende-se ainda, apresentar uma sistematização dos termos gerados a partir de mineração de textos para complementar o processo de indexação manual, visando o aumento do índice de precisão de resposta no processo de busca e recuperação da informação. A Figura 2 ilustra esta questão:

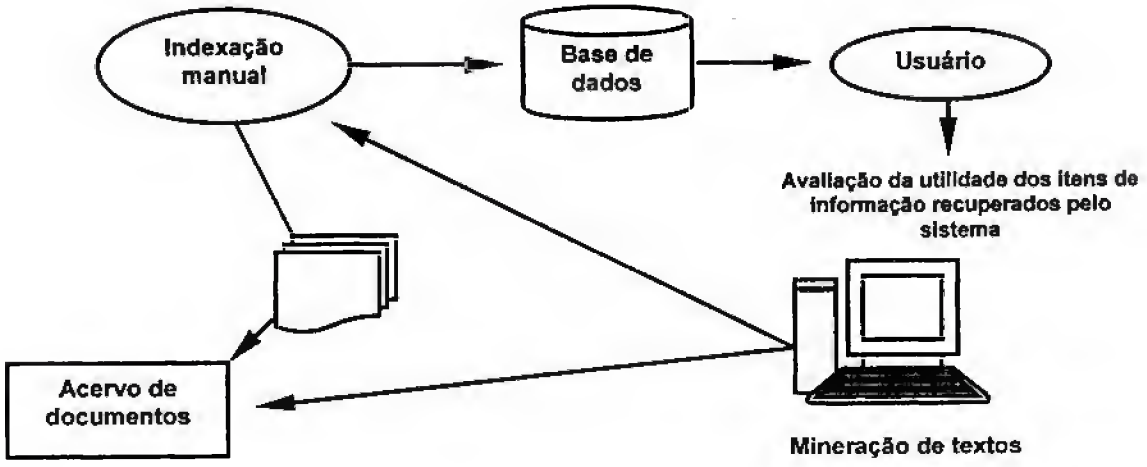


Figura 2 – Posição da mineração de textos no contexto da indexação manual

As três questões que norteiam o problema em estudo são:

A) A mineração de textos aplicada ao processo de busca e recuperação da informação traz ganhos de precisão se comparada à indexação manual?

B) A mineração de textos pode ser empregada como ferramenta complementar no processo de indexação visando o aumento do índice de precisão na recuperação da informação?

C) É possível a construção de uma sistemática de uso de mineração de textos para complementar e aperfeiçoar o processo de indexação visando o aumento do índice de precisão na recuperação da informação?

1.3- Premissas básicas

O problema proposto se assenta nas seguintes premissas:

I. Ganhos no índice de precisão no processo de busca e recuperação da informação poderão ser conseguidos, utilizando-se mineração de textos em associação com a indexação manual; e

II. O julgamento do indexador é preponderante na montagem de uma sistemática de uso dos termos gerados a partir da mineração de textos para auxiliar o processo de indexação manual.

A Figura 3 a seguir, apresenta um diagrama do tipo "espinha de peixe" que traz a problemática proposta pela pesquisa, por meio da reunião dos elementos envolvidos na precisão da resposta entre o processo de busca e recuperação da informação e o reconhecimento da utilidade dos itens de informação recuperados pelo usuário:

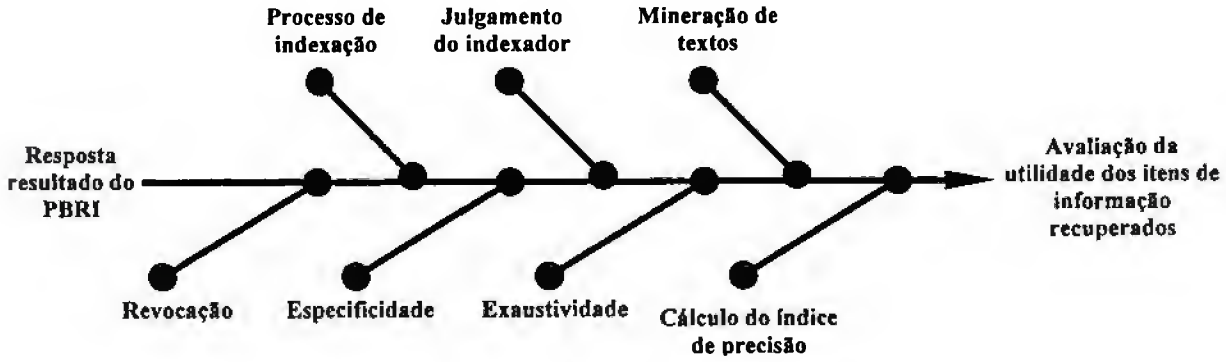


Figura 3 – Fio condutor da resposta à precisão

2- JUSTIFICATIVA

Em inúmeros trabalhos tem-se empreendido esforços na investigação da otimização do processo de busca e recuperação da informação. A consideração das falhas em sistemas de recuperação da informação vem sendo focalizada em sua saturação, ou seja, capacidade insuficiente de dar respostas úteis às formulações das necessidades informacionais dos usuários.

Por certo, a investigação das questões que envolvem a problemática da utilidade da resposta em um processo de busca e recuperação da informação é uma questão legítima de gestão informacional no âmbito da Ciência da Informação, e como tal, exigindo amplas considerações e pesquisas.

O presente estudo se construirá apoiado em pesquisas anteriores no contexto da Ciência da Informação, procurando trazer para a consideração do processo de busca e recuperação da informação o emprego de mineração de textos, auxiliando o indexador e o usuário na busca da melhor resposta para suas demandas.

2.1- Trabalhos afins

Na literatura internacional, cotejada por meio do acesso às bases de dados do *Dissertation Abstracts*, da *Lisa* (em Ciência da Informação) e à base de dados da *Uncover*, todas disponíveis no *Information Ressource Center* da Biblioteca da Embaixada dos EUA em Brasília, os resultados foram semelhantes também à consulta da Base de Teses Brasileiras disponível no Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT. Foi possível localizar a partir de 1998, teses que discutem as falhas nos sistemas de recuperação da informação sem, contudo, relacionarem ferramentas e soluções baseadas na utilização da concepção da mineração de textos. Mesmo assim, relacionamos três teses que se aproximam do escopo do problema da pesquisa ora desenvolvida.

Os trabalhos afins aqui descritos foram divididos em quatro grupos de análise:

A) Os que enfatizam o uso da mineração de textos voltado para o tratamento de dados lingüísticos ou como área emergente de ampla aplicação na gestão do conhecimento;

B) Aqueles que apontam e discutem alternativas às limitações dos sistemas de recuperação da informação comumente utilizados, que influem por sua vez, na precisão da resposta para o usuário;

C) Os que apontam limitações nos sistemas de recuperação da informação atualmente em uso, propondo para saná-las, o emprego de ferramentas destinadas ao aprimoramento do emprego de tesauros e do processo de indexação, utilização da função de crença em bases textuais e tratamento automático de ambigüidades na recuperação da informação; e

D) Aqueles que elucidam as possibilidades de aplicação de mineração de textos no processamento da informação.

O primeiro trabalho do primeiro grupo de trabalhos afins é a dissertação de Fraser (2001), defendida na *Université Laval* no Canadá, sob o título: *Pistes d'exploration pour l'élaboration d'un système formel de montée en abstraction et d'émergence de catégorisations*. Trata da Engenharia lingüística que tem como objetivo fornecer modelos para a concepção e montagem de sistemas informáticos de tratamento de dados lingüísticos. Esta disciplina é utilizada no trabalho de Fraser (2001), no estudo da construção de representações semânticas, aquisição e modelagem de conhecimento por meio de textos, resumos automáticos, tradução automática, mineração de dados e mineração de textos.

O segundo trabalho é a tese de doutoramento de Girju (2002), defendida na *University of Texas*, nos Estados Unidos, sob o título: *Text mining is a rapidly emerging field concerned with the extraction of concepts, relations and implicit knowledge*. Na tese discute-se o aparecimento da mineração de textos como abordagem emergente que diz respeito à extração de conceitos, relações e conhecimento implícito em textos. É proposta a aplicação da mineração de textos, enfatizando a utilização de itens semânticos a fim de descobrir relações implícitas no texto. Demonstra, ainda, a utilidade da mineração de textos em aplicativos avançados no tratamento da linguagem natural.

Na tese intitulada: *Text mining and knowledge discernment: an exploratory investigation*, de autoria de Trybula (1999), defendida na *University of Texas* nos Estados Unidos, a mineração de textos apresenta-se como meio eficaz de identificar padrões escondidos em bases de dados. A tese investigou a metodologia que definiu os passos necessários para a implementação da mineração de textos, bem como a evolução dos seus parâmetros de aplicação à luz da sua funcionalidade em bases de dados. Trybula (1999) testou, ainda, a ferramenta e identificou a evolução do estado da tecnologia de descoberta de conhecimento em textos.

Outro trabalho que merece ser mencionado é a tese de doutoramento de Ong (2004) defendida na *University of Arizona* nos Estados Unidos, sob o título: *Language – and domain – independent knowledge maps: a statistical phrase indexing approach*. No trabalho o autor afirma que a globalização aumentou a necessidade de uso de sistemas multilinguais[1], já que cada domínio, cada base de dados consiste em grande repositório de conhecimento, geralmente capturado em textos. A velocidade com que as informações textuais são produzidas, excede a velocidade que um indivíduo emprega para processá-las, então torna-se necessária a utilização de sistemas automatizados para lidar com esta sobrecarga de informações.

[1] **Sistemas multilinguais** – sistemas que possuem em suas bases de dados documentos em várias línguas.

O autor afirma que ao contrário dos dados inseridos nas bases de dados, um texto que não esteja estruturado não pode ser compreendido de imediato depois de processado por computadores. A tese pretendeu criar uma linguagem e um domínio que aproxima os usuários de mapas hierárquicos do conhecimento, permitindo a compreensão dos conceitos inerentes ao texto – fonte do conhecimento.

A metodologia usada foi o desenvolvimento de um protótipo que analisava a demanda de pesquisa, onde o processamento do conhecimento textual foi realizado por um algorítmo estatístico de indexação de frases aplicado à língua chinesa. Na etapa seguinte, o algorítmo foi expandido para processar múltiplas línguas/domínios e os resultados foram aplicados sucessivamente ao estudo de caso utilizando a estrutura automatizada montada no protótipo, para gerar mapas hierárquicos do conhecimento em coleções chinesas de notícias.

Ong (2004) conclui que uma busca automática é eficiente na criação de mapas do conhecimento e na identificação de conhecimento subjacente, combinando o algorítmo estatístico da extração de frases para representar o conhecimento existente em textos e redes neurais[2] para o agrupamento relacionado aos conceitos existentes no texto. O estudo fornece, também, um conjunto de ferramentas independentes da linguagem do texto para extrair frases de um texto no âmbito do processo de mineração de textos.

Há também a tese de doutoramento de Chung (2004) defendida na *University of Arizona* nos Estados Unidos e intitulada *An automatic text mining for knowledge discovery on the Web*. Segundo o autor, o volume de informações disponíveis observou um acréscimo considerável após o advento da Web, o que superou em muito a capacidade humana de análise de todas as informações

[2] **Redes neurais** – processadores ou *softwares* cuja arquitetura está baseada na estrutura neurológica reticulada do cérebro humano. Estas redes podem processar diversas peças de informação ao mesmo tempo, além de aprender a reconhecer os próprios padrões e programas para solucionar sozinhas problemas parecidos e/ou repetidos (O'Brien, 2000).

disponibilizadas. A sobrecarga de informações está se tornando um problema extremamente complexo, com isso a descoberta de conhecimento na Web tornou-se um grande desafio. A partir deste pressuposto a tese investiga o uso da mineração de textos na descoberta de conhecimento na Web. Para tanto, a mineração de textos englobou cinco etapas:

I. Coleção;
II. Conversão;
III. Extração;
IV. Análise; e
V. Visualização.

Os dados aplicados na pesquisa vieram da Web e posteriormente foram submetidos a cada uma das etapas citadas. A combinação de dados e das técnicas de mineração de textos foram postos para auxiliar a análise humana em diferentes contextos. A questão básica era saber como a descoberta do conhecimento poderia funcionar empregando a estrutura da mineração de textos.

Três investigações empíricas foram conduzidas na área de *business intelligence*[3], que pode ser traduzida como inteligência de negócios:

I. A estrutura foi aplicada na construção de um portal de pesquisa na Área de inteligência de negócios franqueando acesso à pesquisa, sumário de páginas da *Web* e categorização de resultados. Os usuários determinam a pesquisa e avaliam os resultados. Desta forma, a estrutura pôde ser usada na análise e na integração de informações heterogêneas distribuídas em diversas fontes;

II. A estrutura foi desenvolvida para permitir o emprego de dois métodos de pesquisa utilizando o agrupamento. Em termos de precisão e revocação os dois

[3] **Business intelligence** – "análise de decisões baseadas em computador, normalmente realizada online por gerentes e suas equipes. Inclui previsões, análise de alternativa e avaliação de risco e de desempenho (Turban; McLean & Wetherbe, 2004).

métodos apresentaram bons resultados. Os usuários preferiram selecionar os resultados que achavam úteis e de qualidade; e

III. A estrutura classificou as páginas da Web em diferentes categorias de assuntos na abrangência da inteligência de negócios. Após esta etapa, foi proposta uma estrutura que considerava uma relação entre os assuntos encontrados.

Chung (2004) conclui que a estrutura de mineração de textos consegue apoiar o usuário na identificação de informações mais precisas em um grande volume de informações textuais.

Finalmente, o único trabalho encontrado na literatura internacional até 2004 que associa claramente o processo de busca e recuperação da informação com a mineração de textos foi a tese de doutoramento de Goldman (1998), defendida na *University of California* nos Estados Unidos, intitulada: *A digital filter model for data mining of text documents (databases, computacional linguistics)*. A partir da consideração da grande proliferação e profusão de dados em diferentes suportes, objetos, imagens, áudio, vídeo e textos de formatos livres, Goldman discute a necessidade cada vez mais premente de se retirar conhecimento deste emaranhado de dados, apontando a descoberta de conhecimento em base de dados como a 'nova' ciência responsável por isto. O trabalho apresenta, então, um novo modelo e uma nova arquitetura: o '*digital filter*' ou filtro digital que é baseado nas premissas da mineração de dados, da recuperação da informação e da lingüística computacional. A metodologia do filtro digital explora as informações no que diz respeito a distribuição de palavras de documentos textuais que permitem a identificação de conhecimento, quando os mesmos são submetidos à abordagem do filtro digital. Este modelo seria capaz de encontrar palavras atípicas, frases ou outras estruturas lingüísticas de acordo com determinados contextos.

São apresentados resultados da aplicação do filtro digital em textos de coleções específicas, tais como: artigos, reportagens de jornais, dados de textos livres de bases de dados relacionais, decisões jurídicas e programação de cinema. Goldman explorou ainda, na tese, dois casos que levaram a descobertas significativas. O primeiro foi a respeito de uma coleção de prontuários médicos com diagnóstico de câncer toráxico e o segundo referiu-se a relatórios de atividades sísmicas. Por fim, a tese conclui que o filtro digital pode ser um modelo aplicável na recuperação da informação, como um indexador e como identificador de problemas relacionados á detecção de plágio.

Na literatura brasileira em Ciência da Informação, cotejada a partir do acesso à base de dados das Teses Brasileiras do Instituto Brasileiro de Informação em Ciência e Tecnologia – IBICT, verificaram-se contribuições na abordagem do problema ora proposto, entretanto, os trabalhos são focados basicamente nas questões relativas às falhas dos sistemas de recuperação da informação, sem, na maioria dos casos, trazerem soluções que envolvessem usuários, profissionais da informação e aplicação de instrumentos de Gestão do Conhecimento baseados em mineração de textos no processo de busca e recuperação da informação.

No âmbito das teses brasileiras, o trabalho de doutoramento de Araújo (1994) aborda a problemática das falhas dos sistemas de recuperação da informação – SRI em atender as expectativas dos usuários. A autora questiona se para corrigir tais falhas, não será necessário um novo modelo ou uma nova abordagem teórico-conceitual que estude os SRI como são de fato e não como se imagina que poderiam ser. A tese firma como objetivo responder aos questionamentos: por que vêm falhando os sistemas, por que vêm falhando os SRI?

Desta forma, Araújo (1994) formula sua hipótese de pesquisa, apresentando dois considerandos. O primeiro, de que a explosão exponencial da

informação ocorre desde o século XVII em função da revolução científica e do periódico científico; e o segundo de que os sistemas de informação construídos pelo homem não mais cumprem seus objetivos de organizar e disseminar a massa gigantesca de informações devido ao seu grande crescimento. Em seguida, é apresentada a seguinte hipótese:

- O sistema de informação, enquanto sistema artificial/social, está atingindo o seu limite de crescimento, saturando-se, exigindo, assim, uma inversão em seu crescimento exponencial. A reversão do sistema de informação a tamanhos menores, mais adequados, é condição necessária (mas não suficiente) à sua sobrevivência enquanto sistema social.

Em outra pesquisa, os sistemas de recuperação são vistos como sistemas de 'redução da informação'. O objetivo da dissertação de Pereira (1994) está em discutir o fato de que os processos de representação temática do conteúdo dos documentos, especificamente o processo de indexação, "reduz, fecha e rotula" a informação, criando uma série de problemas que irão desembocar na 'redução da informação'. A autora propõe os seguintes considerandos para a formulação da sua hipótese:

- Os sistemas apresentam uma única visão da realidade, para então propor modelos em substituição a esta realidade;
- Os SRI são sistemas que, ao processar e disseminar documentos, dão acesso a um universo de conhecimento representado por esses documentos;
- O subsistema representação é mecanismo-chave de acesso a esse universo de conhecimento, gerando os modelos operacionais daqueles documentos; e

- As representações do universo de conhecimento expressas pelos documentos e as expressas pelos modelos gerados pelo sistema (indexação) são disjuntos em suas visões de mundo.

É a seguinte a hipótese apresentada por Pereira (1994):

- Os sistemas de recuperação da informação – SRI, ao modelarem o universo de conhecimento a que se referem, desviam, escondem e mutilam esse mesmo universo que se propõem a revelar.

Segundo Pereira (1994), ao impingir uma visão de mundo unidirecionado, os sistemas de recuperação da informação desviam de outras percepções possíveis todos os demais olhares. Escondem as relações intra e inter documentos, impedindo que essas relações sejam reveladas, explicitadas. Mutilam a informação, tomando a parte pelo todo, engendrando, um simulacro daquele todo.

Tanto o trabalho de Araújo (1994) quanto o de Pereira (1994) guardam semelhanças ao discutirem as falhas do SRI que influem na precisão da resposta para o usuário. Chama-se a atenção, no caso da primeira pesquisa, para a saturação desses dois sistemas, e na segunda, para uma questão mais pontual, que será um dos elementos-chave na pesquisa ora empreendida: a obliteração dos resultados de um SRI, justamente naquilo que deveria ser o instrumento de aumento da precisão da resposta no processo de busca e recuperação da informação, ou seja, a representação temática do conteúdo dos documentos por meio do processo de indexação.

A seguir, passamos a relacionar o segundo grupo de trabalhos os quais, também a partir da perspectiva da recuperação da informação, aproximam-se como temas correlatos ao estudo da precisão no processo de busca e recuperação da informação.

Na dissertação de Zavitoski (2001), intitulada Exploração do uso do tesauro como instrumento de recuperação da informação, diferentemente dos questionamentos formulados por Pereira (1994), constata-se que a representação temática do conteúdo dos documentos é um instrumento plausível para a recuperação da informação, desde que garantidas as relações de equivalência entre as linguagens dos documentos, usuários e profissionais da área. Assim sendo, o tesauro é apresentado como instrumento de uso efetivo destes profissionais (técnicos de pesquisa) na recuperação da informação.

Dados estes pressupostos, a autora apresentada suas hipóteses de trabalho:

- Os tesauros, como vêm sendo desenvolvidos atualmente, independentemente de seguirem rigorosamente os padrões consagrados internacionalmente, dificultam o acesso do usuário final à informação, na medida em que não integram a linguagem do seu usuário, comprometendo a recuperação da informação; e
- A recuperação da informação é comprometida, também devido à inconsistência da indexação, quando os tesauros se restringem somente a estruturar a área de especialidade e a controlar o vocabulário, e não contemplam as remissivas que possibilitam ao indexador efetuar a compatibilização das linguagens do autor, com as linguagens de especialidades.

Zavitoski (2001) pretende, portanto, rever o papel do tesauro no auxílio da estratégia de busca na recuperação da informação.

A pesquisa de Ramos (1999) é voltada para a qualidade da recuperação da informação em bases textuais, com o emprego da função de crença[4] em lugar do

[4] **Função de crença** - Ferramenta usada na área de Inteligência Artificial e definida em termos de domínios discretos finitos conhecidos como quadro de discernimento, a teoria da função de crença tem sido largamente

modelo booleano[5]. Neste caso, há o emprego de um instrumento, assim como no trabalho de Zavitoski, centrando esforços especificamente no mecanismo de busca em bases textuais. Na dissertação são enumeradas hipóteses a fim de demonstrar matematica e experimentalmente que o uso do mecanismo de busca de função de crença, se materializa com maior grau de qualidade no processo de recuperação da informação do que o de busca booleana, contribuindo, assim, para a satisfação dos usuários.

O trabalho de autoria de Orrico (2001), intitulado: Binômio lingüística – ciência da informação: abordagem teórica para elaboração de metafiltro de recuperação da informação, apresenta como objetivo a proposição de um modelo lingüístico para tornar mais eficaz a recuperação da informação. A autora relata que o estudo está centrado na Ciência da Informação, especificamente na área relativa aos processos de busca e recuperação da informação apoiados em conceitos teórico-metodológicos da Lingüística.

A tese apresenta a seguinte premissa:

- A fundamentação teórica da Lingüística serve de suporte ao cientista da informação para que ele, ao melhor depreender o significado construído nas comunicações em geral, possa intermediar uma recuperação eficaz.

Desta forma, a tese defendida no trabalho é que o metassistema de representação – que se origina nas representações formuladas pela comunidade diretamente envolvida e que se delineia a partir de representações bem sucedidas nos ambientes físico e cultural – pode oferecer indicadores para o desenvolvimento de filtros para a recuperação da informação, na medida que tal

---

empregada a fim de representar e atualizar o conhecimento impreciso e realizar o raciocínio evidencial automatizado (Ramos, 1999).

[5] **Modelo booleano** - Baseia-se no sistema dicotômico onde somente existem dois estados que se excluem mutuamente. Os termos modelo booleano, lógica booleana, álgebra booleana ou lógica matemática são utilizados como sinônimos (Daghlian, 1995).

metassistema leva em conta parâmetros de relevância e pertinência na construção dessas ferramentas de busca (Orrico, 2001).

Por conta de sua tese central, a autora desdobra o problema em duas vertentes:

I. os interlocutores, ou seja, os que divulgam o conteúdo informacional e os que buscam tais conteúdos; e

II. o significado como um processo de construção de interação entre os interlocutores.

Assim sendo, conceitos como precisão e especificidade são discutidos ao longo do trabalho.

A tese de autoria de Medeiros (1999) intitulada: Tratamento automático de ambigüidades na recuperação da informação encaixa-se, também, no estudo da recuperação da informação, apresentando um instrumento para a redução da ambigüidade em sistemas de recuperação da informação. Basicamente, a autora trata da extração de informações contidas em textos completos por meio de métodos de tratamento automático da linguagem natural. O enfoque da aplicação deste tratamento parte dos fenômenos lingüísticos, notadamente da ocorrência da ambigüidade[6]. Diante disto, foram formuladas duas perguntas geradoras pela autora:

- As ambigüidades que ocorrem em textos científicos e técnicos em língua portuguesa e que interferem na recuperação da informação podem ser solucionadas por meio do tratamento da linguagem natural?

[6] **Ambigüidade** - Ocorre quando palavras ou frases podem gerar mais de uma interpretação de seu significado (Medeiros, 1999).

- Como introduzir informações semânticas num sistema de tratamento automático da linguagem natural, de maneira a possibilitar a solução destas ambigüidades?

Neste momento, pode-se afirmar que o trabalho de Medeiros (1999), assim como o de Orrico (2001), são as investigações que mais se aproximam da proposta de pesquisa ora apresentada. As diferenças básicas são: promoção da precisão na recuperação da informação em Orrico e Medeiros e verificação do índice de precisão na recuperação da informação por meio da comparação entre a mineração de textos e a indexação manual no trabalho ora proposto. Em Medeiros (1999) o instrumento proposto capaz de promover o aumento da precisão está no tratamento automático de ambigüidades, por meio da criação de bases de conhecimentos morfológicos, sintáticos e semânticos em língua portuguesa.

Finalmente na consideração de trabalhos semelhantes, apresentamos as pesquisas que enfocam e demonstram as possibilidades da mineração de textos.

A dissertação intitulada Um estudo sobre agrupamento de documentos textuais em processamento de informações não estruturadas usando técnicas de *clustering* de autoria de Wives (1999) apresenta como objetivo o uso de técnicas de agrupamento de objetos (informações) textuais a fim de facilitar o acesso à informação. Com isso, a tarefa de identificação de documentos úteis para o usuário em uma grande coleção de documentos é facilitada pelo agrupamento em torno de um mesmo assunto.

Segundo o autor, a motivação da pesquisa nasceu da consideração da sobrecarga da informação, ampliada sensivelmente pelo advento da Internet, onde o grande volume de dados não favorece a recuperação de uma 'boa' informação.

O método de busca e recuperação por termo ou palavra-chave origina inúmeros problemas que são, segundo o autor, inerentes à linguagem utilizada.

Pertencem a uma classe de problemas do vocabulário oriundos da própria natureza ambígua da linguagem. Desta maneira, o vocabulário utilizado pelo sistema pode diferir daquele empregado por quem realiza a busca. Wives (1999) lembra, ainda, que a identificação da 'boa' informação demanda horas diante de uma ferramenta de busca na Internet, e muitas vezes vem associada a muitas outras distribuídas em vários documentos, onde é necessário analisar seu conteúdo para filtrar e extrair o que é de fato útil ao usuário.

Diante disto, a mineração de textos apresenta-se como uma ferramenta capaz de sumarizar um conjunto de documentos em agrupamentos, apresentando-os sob forma de gráficos indicativos das relações semânticas dos termos que os compõem. Assim, o usuário obtém uma idéia mais clara do assunto de que trata a coleção de páginas, sem que para isso tenha que lê-las uma a uma.

O usuário pode ainda, segundo Wives (1999), refazer a consulta aprofundando-se no assunto ou buscando novos conhecimentos por meio de correlações de agrupamentos de informação.

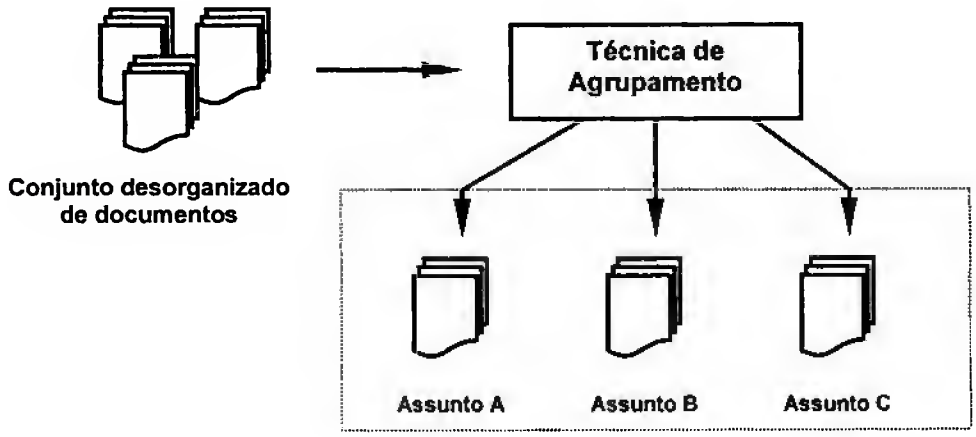


Figura 4 – Objetivo do agrupamento de informações textuais (Fonte: Wives, 1999)

O último trabalho do levantamento bibliográfico de trabalhos afins intitula-se Descoberta de conhecimento com o uso de *text mining*: cruzando o abismo de Moore, de autoria de Silva (2002). O objetivo da dissertação foi:

- Estudar e propor alternativas para a travessia do chamado abismo de Moore[7] pela tecnologia de descoberta de conhecimento em textos[8];

Em complementação a este objetivo são apresentadas três finalidades:

- Propor uma forma de trabalho para extração do conhecimento a partir de bases textuais;
- Explorar a possibilidade de uso de uma metodologia de descoberta de conhecimento em base de dados para desenvolver aplicações em descoberta de conhecimento em textos;
- Mostrar a possibilidade de uso efetivo da mineração de texto por meio de um estudo de caso real.

2.2- Conclusão e proposta do autor

As tendências percebidas nos trabalhos afins apóiam a afirmação de que a comparação da utilidade da ferramenta de mineração de textos em relação à indexação manual, pode ser uma efetiva contribuição na verificação do aproveitamento de novas tecnologias da informação na otimização do processo de busca e recuperação da informação.

---

[7] **Abismo de Moore** - a partir de um modelo de descrição do comportamento dos consumidores de tecnologias em áreas emergentes, Geofrey Moore apresenta cinco tipos de usuários: inovadores; adeptos iniciais; maioria inicial; maioria tardia e os retardatários. Diante disto, Moore constata que há um "abismo" entre os adeptos iniciais e a maioria inicial e que este "abismo" é onde a maioria das organizações falha por não dispor de instrumentos efetivos de *marketing* para lidar com a problemática (Silva, 2002).

[8] Na literatura os termos: descoberta de conhecimento em textos e mineração de textos são comumente usados como sinônimos, o mesmo acontece, com menor freqüência, com a descoberta de conhecimento em bases de dados e mineração de dados (Benoît, 2002).

O passo seguinte da investigação é a comparação de um processo tradicional e consagrado na representação do conteúdo dos documentos, como é o caso do processo de indexação, com a mineração de textos. Para viabilizar esta comparação, optou-se pelo uso do índice de precisão, que é uma medida objetiva e consagrada desde a sua proposta por Cleverdon (1962), além de ser capaz de mostrar as diferenças de desempenho entre a indexação e a mineração de textos na recuperação da informação.

Sobre a factibilidade de incluir na problemática da recuperação da informação a mineração de textos, pode-se citar, dentre os autores estudados, a percepção de uma lacuna: a utilização da mineração de textos voltada para a resolução das falhas no processo de busca e recuperação da informação. Nesta associação há de fato uma possibilidade clara de uso da ferramenta na melhoria de performance destes sistemas, todavia a cautela aponta na direção da utilização da mineração de textos para auxiliar o processo de indexação, já que a grande potencialidade da ferramenta está na sua capacidade de sumarizar grandes conjuntos de documentos sob a forma de agrupamentos, apresentando-os sob a forma de listas de palavras que mais ocorrem por documento ou por resultado de pesquisas (conjuntos de documentos), e em alguns casos com gráficos indicativos das relações semânticas entre os termos. Assim, a possibilidade de extrair de uma montanha de textos, informação útil às demandas, torna-se um efetivo instrumento de gestão em bases textuais. Contudo, a informação útil não é dada de forma automática com a mineração de textos, mas por meio da interpretação que for dada aos resultados obtidos.

Assim sendo, o aumento do índice de precisão da resposta obtida do processo de busca e recuperação da informação, poderá ser alcançado a partir de dois Fatores Críticos de Sucesso (FCS)[9]:

[9] **Fatores Críticos de Sucesso (FCS)** – características, condições ou variáveis críticas para o sucesso (atingimento dos objetivos) em um dado processo ou até mesmo em uma organização (Rockart, 1979).

I- Aplicação da mineração de textos; e

II- Integração dos resultados obtidos com a mineração de textos à indexação.

O Quadro a seguir, associa os fatores críticos de sucesso aos seus objetivos – chave, a fim de torná-los mais claros:

Quadro 1 – FCS na gestão da precisão no processo de busca e recuperação da informação

FATORES CRÍTICOS DE SUCESSO (FCS)	*OBJETIVOS – CHAVE*
Aplicação da mineração de textos	Auxiliar o processo de indexação
Integração dos resultados obtidos com a mineração de textos à indexação	Montagem de uma sistemática de uso da mineração de textos no processo de indexação, com vistas ao aumento do índice de precisão no processo de busca e recuperação da informação

Desta forma, a proposição do estudo comparado entre a mineração de textos e a indexação manual, por meio da análise do índice de precisão de resposta no processo de busca e recuperação da informação, pretende contribuir efetivamente no âmbito da Ciência da Informação para o estabelecimento de estratégias de uso da mineração de textos na melhoria contínua da resposta nestes sistemas, além de verificar, com clareza, quais os ganhos que a mineração de textos pode trazer em relação ao processo de indexação na recuperação da informação.

3- OBJETIVOS DA PESQUISA

3.1- Objetivo geral

Comparar a utilidade da ferramenta de mineração de textos com a lista de palavras-chave utilizadas na indexação manual por bibliotecários da Caixa Econômica Federal, verificando a variação no índice de precisão no processo de busca e recuperação da informação na base de dados do Infohab.

3.2- Objetivos específicos

A) Avaliar se a recuperação da informação pela aplicação da ferramenta de mineração de textos traz ganho no índice de precisão, se comparada com a lista de palavras-chave utilizadas na indexação manual na base de dados do Infohab.

B) Verificar a viabilidade de se utilizar os termos resultantes do emprego de ferramenta de mineração de textos, para enriquecer a lista de palavras-chave usada no Infohab, objetivando aprimorar o trabalho de indexação manual em relação ao índice de precisão de resposta no processo de busca e recuperação da informação.

C) Propor uma sistemática de uso dos termos gerados a partir da mineração de textos, que apóie o processo de indexação manual, visando o aumento do índice de precisão de resposta no processo de busca e recuperação da informação.

4- REVISÃO DE LITERATURA

Considerando o tema: estudo comparado entre a mineração de textos e a indexação manual, por meio da análise do índice de precisão de resposta no processo de busca e recuperação da informação: o caso do Centro de Referência e Informação em Habitação – Infohab e o problema a ser investigado, este capítulo foi estruturado com as seguintes seções: 4.1- O processo de indexação: 4.1.1- Indexação manual e indexação automática, 4.1.2- Análise documentária e a representação do conteúdo dos documentos, 4.1.3- Linguagens de indexação, 4.1.4- Coerência e qualidade da indexação, 4.1.5- Conclusão; 4.2- A mineração de textos: 4.2.1- A mineração de textos e a mineração de dados, 4.2.2- Tipologia da mineração de textos, 4.2.3- Conclusão; 4.3- O processo de busca e recuperação da informação: 4.3.1- Sistemas de recuperação da informação, 4.3.2- Conclusão; e 4.4- Precisão: 4.4.1- Conceitos e índice de precisão, 4.4.2- Gestão da precisão, 4.4.3- A precisão no processo de busca e recuperação da informação, 4.4.4- A mineração de textos e o índice de precisão, 4.4.5- Conclusão.

4.1- O processo de indexação

A análise, descrição e representação temática do conteúdo de documentos são capitais ao se definir a indexação e o seu processo. Outros itens concorrem para a definição: a base de dados onde a representação do conteúdo dos documentos estará armazenada e o vocabulário do sistema, ou seja, o vocabulário controlado onde os termos empregados na indexação poderão formar, desde que exista uma estrutura semântica, um tesauro. O vocabulário do sistema, por sua vez, influenciará as estratégias de busca doravante utilizadas na busca e recuperação da informação nas bases de dados.

A Figura 5 mostra a posição da indexação na recuperação da informação.

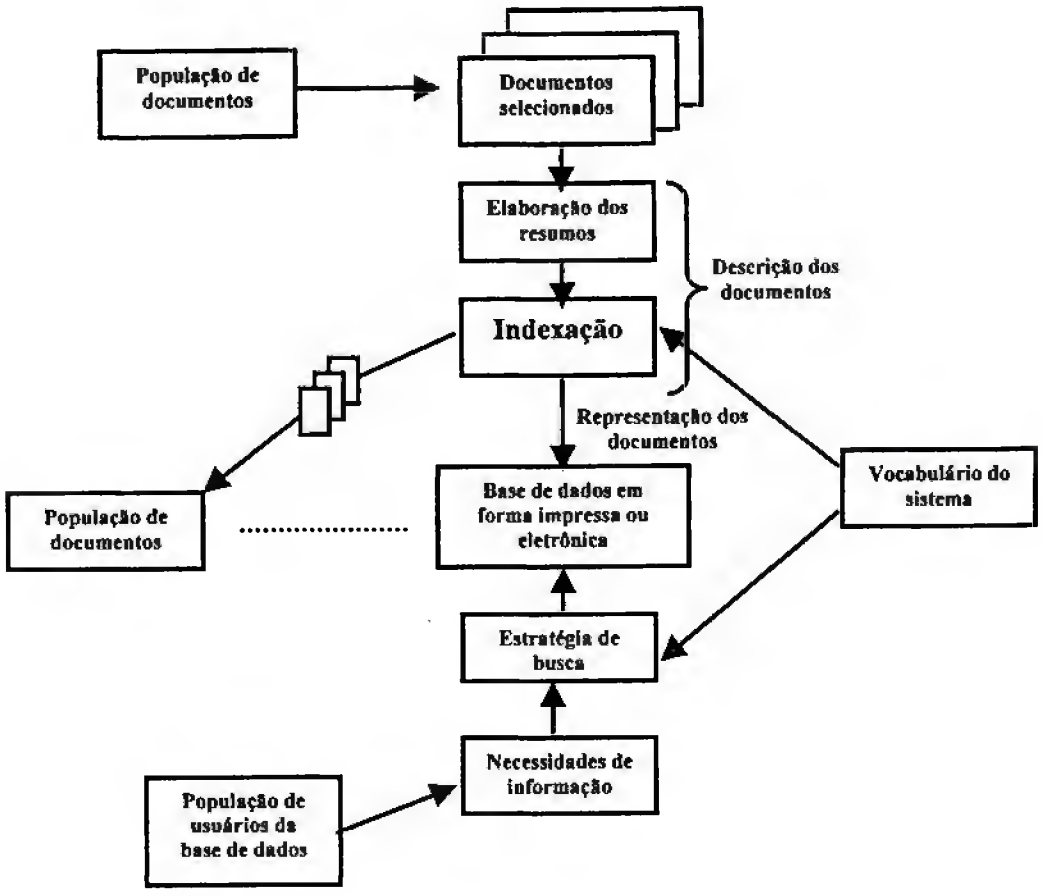


Figura 5 – Elaboração de índices e resumos na recuperação da informação (Fonte: Lancaster, 1998)

Ao pensar a indexação no âmbito maior das funções de um sistema de recuperação de informação, tal como proposto por Cianconi (1990), pode-se verificar a indexação ligada à administração da informação e mais especificamente no contexto da sua organização (Figura 6):

Figura 6 – A indexação no âmbito das funções de um sistema de recuperação da informação
(Fonte: adaptado de Cianconi, 1990)

A indexação pode ser definida como: "tradução de um documento em termos documentários, isto é, em descritores, cabeçalhos de assunto, termos-chave, que têm por função expressar o conteúdo do documento" (Cintra, 1983) ou como o processo de atribuir termos ou códigos de indexação a um registro ou documento, termos ou códigos esses que serão úteis posteriormente na recuperação da informação (Rowley, 2002).

Para Robredo & Cunha (1986), a indexação é o processo pelo qual se identificam os conceitos de que trata o documento, expressando-os na terminologia usada pelo autor (linguagem natural) ou com o apoio de vocábulos ou termos de significação unívoca ou, ainda, por meio de códigos (linguagens documentárias, descritores, sistemas de classificação, etc.).

A conversão da linguagem na qual um documento foi redigido para uma linguagem documentária vai envolver uma leitura analítica do documento por parte do indexador, a fim de identificar e selecionar palavras-chave (indexadores) que possam representar de forma fidedigna o seu conteúdo. Segundo Cintra (1983), a leitura realizada pelo indexador tem como propósito selecionar descritores que sejam compatíveis com uma determinada linguagem documentária, para que o documento seja posteriormente recuperado.

Na definição oferecida por Baranow (1983), a indexação é vista como um processo analítico onde aparece decomposta em duas etapas distintas: descrição e representação. A primeira etapa refere-se a identificação, seleção e análise dos conceitos que de fato representam o conteúdo de um dado documento e a segunda a representação desses conceitos através de descritores (termos) compatíveis com os do sistema de recuperação. Assim sendo, as duas etapas podem ser resumidas como a análise de conteúdo e a escolha dos conceitos que representam esse conteúdo.

Há ainda a definição proposta por Slype (1991): "a indexação é a atividade que consiste em representar o conteúdo de um documento ou de uma consulta de modo analítico, ou seja, enumerando conceitos e/ou palavras".

No momento em que se percebe a existência de etapas diferenciadas na tarefa da indexação, já se pode pensar em um processo de indexação. Assim sendo, Rowley (1988) distingue três estágios no processo de indexação:

A distinção das etapas do processo de indexação propostas por Rowley (1988), coincide com os argumentos apresentados por Baranow (1983), de que a indexação deve ser vista como um processo analítico, onde a descrição e a

representação estão presentes. Desta forma, a análise acaba sendo um fator preponderante nas duas visões.

Cabe ainda verificarmos as argumentações levantadas por Wellisch (1995) que relatam que a maioria das abordagens de como elaborar índices, valia-se de um modelo já programado que consistia em identificar palavras e nomes "importantes", escrevê-los em fichas, acrescentar número das páginas onde aparecem e então colocá-las em ordem alfabética, preparando também uma lista de entradas.

O autor relembra que o advento do computador pessoal apenas facilitou a elaboração de índices, mas a base continuou sendo exatamente a mesma de antes. Como exemplo Wellisch (1995) cita o artigo publicado em uma revista especializada em informática, que foi apelidado de "indexação espontânea": para indexar um livro é necessário seguir os seguintes passos: "1- Remover toda a pontuação, letras maiúsculas, apóstrofes, etc e colocar todas as palavras do livro em uma lista em separado; 2- Eliminar todas as palavras repetidas; 3- Organizar as palavras alfabeticamente; 4- Remover as palavras não significativas (e, para, então, etc); e 5- Indicar as páginas onde as palavras aparecem no livro." (Pountain, 1987 *apud* Wellisch, 1995).

Esta passagem é avaliada por Wellisch como sendo uma caricatura do processo de indexação, pois não considera todas as questões semânticas envolvidas, tais como sinonímia, sintaxe, palavras homógrafas, etc. O processo de indexação seria mais complexo, já que envolve um razoável esforço intelectual, utiliza linguagem específica, às vezes artificial, e em muitos casos aplica a intuição do indexador não podendo, então, se prender a regras fixas.

De toda forma, Wellisch (1995) relaciona itens que considera importantes no processo de indexação para o indexador:

1- Identificar tópicos e assuntos dentro do texto que venham ao encontro das necessidades dos possíveis usuários daquele índice;

2- Separar tópicos que contenham informações diferentes daquelas necessárias aos usuários;

3- Excluir tópicos que contenham informações diferentes daquelas necessárias aos usuários;

4- Analisar conceitos, tópicos e assuntos considerados importantes, a fim de que seja providenciada uma lista de entradas para as mesmas;

5- Produzir cabeçalhos que empreguem terminologia usada no documento, considerando também sinônimos ou termos equivalentes para auxiliar o processo de recuperação da informação;

6- Certificar-se de que o cabeçalho utilizado é apropriado para as necessidades dos usuários e que vai apoiá-los na: a) recuperação rápida de alguma informação contida em um documento; b) identificação rápida da presença ou ausência de informação em um documento; e c) identificação de documentos em uma coleção;

7- Agrupar referências similares nos tópicos selecionados no documento;

8- Combinar cabeçalhos e sub-cabeçalhos em um cabeçalho de multi-níveis que seja coerente;

9- Indicar interface entre conceitos e tópicos;

10- Chamar a atenção do usuário para os termos do tipo "ver" e "ver também"; e

11- Colocar todos os cabeçalhos em uma ordem, sendo que a mais usual é a alfabética.

A indexação como um processo de representação do conteúdo dos documentos, é um elemento fundamental para o processo de busca e recuperação da informação. O armazenamento da informação, só pode ser realizado com efetividade, se a indexação for feita de modo satisfatório, ou seja, representando com fidedignidade o conteúdo dos documentos. A exaustividade é a primeira dimensão da indexação, que segundo Lancaster (1998), corresponde a uma representação exaustiva do conteúdo temático dos documentos, de acordo com dois níveis: I. exaustiva – quanto mais termos forem incluídos no processo de indexação e II. seletiva – quanto menos termos forem incluídos. Há ainda o

conceito de especificidade que se refere a segunda dimensão da indexação, quando um documento deve ser indexado com o termo mais específico que o abranja totalmente.

Para Robredo & Cunha (1986), a indexação pode ser realizada em níveis diferentes, dentre os quais são destacados a **categorização**, que consiste em reconhecer no documento um aspecto dominante, segundo uma certa subdivisão por assuntos; a **indexação superficial**, que permite obter os conceitos principais tratados no documento; e a **indexação profunda**, que consiste em obter todos os conceitos considerados fundamentais.

Há também a indexação de assuntos que segundo Lancaster (1993), é uma expressão usada de modo impreciso, pois refere-se à representação do conteúdo temático de partes de itens bibliográficos completos, como o índice no final de um livro. O autor afirma que a "distinção entre catalogação de assuntos e indexação de assuntos, uma delas referindo-se a itens bibliográficos completos e a outra a partes de itens, é artificial, enganosa e incongruente".

A indexação, seja manual, automática ou mista, influi diretamente, na tarefa de recuperação da informação. Se os descritores selecionados para representar o conteúdo de um dado documento não forem coerentes, certamente não será recuperado com facilidade, fato que comprometerá o processo como um todo.

Mais do que partes do sistema de recuperação da informação, o armazenamento, a recuperação e sobretudo a indexação, são os seus fatores críticos de sucesso.

4.1.1- Indexação manual e indexação automática

A indexação manual ou intelectual consiste na atribuição de termos de indexação ou códigos de indexação realizada por um ser humano. Estes termos serão selecionados e atribuídos por indexadores com base no julgamento

subjetivo realizado com que fazem acerca do conteúdo do documento, ou escolhem termos que tenham probabilidade de virem a ser procurados por um usuários no futuro (Rowley, 2002).

A Indexação manual pode também ser conceituada como a tradução de um documento em termos documentários (descritores, cabeçalhos de assunto e termos-chave) sem o auxílio da atribuição automática de termos ou extração automática de termos; é a indexação sem o auxílio de computadores.

A indexação automática é qualquer procedimento que permita identificar e selecionar os termos que representam o conteúdo dos documentos sem a intervenção direta do documentalista. No processo de indexação automática, um algoritmo (ou seja, um conjunto de operações elementares, organizadas logicamente) realiza, em certa medida, o trabalho do indexador no processo de escolha dos termos significativos (Robredo & Cunha, 1986).

A indexação automática pode também ser conceituada como a indexação realizada com o apoio de computadores que selecionam, por meio de um conjunto de instruções programadas previamente, os termos que mais ocorrem em um documento.

A indexação automática pode ser dividida em dois tipos, segundo Lancaster (1998):

- **Indexação por extração automática:** a indexação por extração consiste na retirada do texto de palavras que serão usadas para representar o seu conteúdo. Indexadores responsáveis por este trabalho deverão selecionar termos e/ou expressões que existem no texto para representar o seu conteúdo temático. Nesta tarefa são naturalmente escolhidas palavras que com maior freqüência ocorrem no texto, sua posição no título ou no resumo e o contexto em que aparecem. Da

mesma maneira, se o texto encontrar-se na forma eletrônica, o computador pode ser programado para selecionar e retirar com precisão absoluta, palavras e/ou expressões que mais ocorrem no texto, identificar sua posição e seu contexto.

- **Indexação por atribuição automática:** este tipo de indexação é realizada pelo computador, desenvolvendo para cada termo atribuído uma espécie de 'perfil', ou seja, um conjunto de palavras e/ou expressões que freqüentemente ocorrem nos documentos e que seriam atribuídas por indexadores humanos. Como exemplo, o autor propõe o perfil para o termo chuva ácida que incluiria no rol de termos: precipitação ácida, poluição atmosférica, dióxido de enxofre, etc.

A discussão em torno do processo de indexação automática também tem incorporado contribuições importantes, cabendo ressaltar aqui, a contribuição que a Lingüística tem trazido desde o início da década de 80, inclusive influenciando a questão da indexação automática. Desta forma, é importante ressaltar a observação que Baranow (1983) e Baranow (1976) oferece sobre o tema: "na prática a indexação vem sendo influenciada intensamente pela automação, enquanto que em nível teórico se critica a falta de uma *Teoria Integrada da Indexação.* Uma teoria desse tipo também teria de esclarecer explicitamente, as relações entre Lingüística e a indexação e procurar incorporar em seu bojo, aquelas partes da Lingüística que lhe são relevantes. Enquanto isso não se tornar realidade, continuamos com as soluções *ad hoc* lingüísticas, para resolver problemas de indexação".

Desde 1980 pesquisas vem sendo desenvolvidas associando a indexação automática com as noções universais de tema-rema ou tópico-comentário da Lingüística. Sobre isto Baranow (1983) afirma que ao elaborar as noções tema-rema, pressupõe-se a existência de uma hierarquia informacional na frase, no período e no texto. Assim, o tema pode ser apreendido por meio da análise da base temática do item de informação em questão. O rema é o que se acrescenta

ao já conhecido, daí a possibilidade de as unidades de um texto se apresentarem hierarquicamente ordenadas a partir da noção de tema-rema.

Anderson & Pérez-Carballo (2001) em seu trabalho sobre a natureza da indexação parte 1, colocam que para a busca de informações, textos ou documentos em um sistema de informação, é necessário que os termos estejam descritos e indexados. A descrição requer alguns tipos de análise, dos quais dois são destacados pelos autores: a análise humana e a análise algorítmica realizada por computadores.

A análise humana parte do exame de documentos e textos para considerar o contexto que representam, além de suas características. A análise automática identifica e compara os componentes do texto – os símbolos que formam o texto – às vezes consultando dados lexicais ou tesauro, a fim de caracterizar conjuntos de componentes textuais, às vezes aplicando indexação sintática para identificar grandes unidades do texto e às vezes calculando atributos para os componentes do texto e documentos, baseados em dados disponíveis.

Os dois tipos de análises, segundo os autores, são empregados na indexação e análise de textos e é comum ouvir dizer que os resultados de ambos são bem diferentes quando se trata da recuperação da informação propriamente dita. Todavia, os usuários registram que as duas abordagens têm mais ou menos o mesmo valor, fato confirmado no trabalho de Bastos (1984), onde a autora conclui que "a indexação automática aplicada aos títulos e resumos dos artigos do periódico Ciência da Informação, entre 1972 e 1983, identificou de maneira equivalente à indexação manual, os descritores que caracterizam a base de dados formada com os referidos artigos".

O ideal é que um sistema de recuperação da informação ofereça as duas abordagens de análise/indexação da informação: tanto a humana quanto a

automática, tornando o processo de busca e recuperação da informação mais exaustivo e eficiente e, como conseqüência, os resultados mais relevantes.

Ainda segundo Anderson & Pérez-Carballo (2001) em seu trabalho sobre a natureza da indexação parte 2, na indexação automática o foco é o método automático utilizado como base no contexto da recuperação da informação. Opções de busca e técnicas, os métodos usados para criar buscas efetivas, agregação de valor aos termos de busca, buscas específicas, utilização de 'truncagem' de termos ou de combinações booleanas e operadores, tudo isso é considerado parte da sintaxe de busca e, portanto, não pode ser abordado pelo método automático de indexação, que se restringe à linguagem textual.

Os autores argumentam ainda que, em se tratando de indexação de materiais especiais como som e imagem, por exemplo, a indexação automática apenas engatinha, uma vez que sua base é a linguagem contida nos textos. O Altavista é uma ferramenta de busca na *Web* que utiliza a indexação automática de imagens desde 1998, tentando encontrar imagens que sejam visualmente similares ao comando de busca determinado pelo usuário. Imagem 'visualmente' similar não é a mesma coisa que imagem 'conceitualmente' similar. Assim sendo, os resultados quase sempre parecem estar baseados na cor e na estampa e não em um detalhe particular da imagem.

4.1.2- Análise documentária e a representação do conteúdo dos documentos

A representação do conteúdo dos documentos está na base do conceito de indexação, bem como de conceitos relacionados, tais como descritor, linguagem de indexação e termo de indexação. No caso da indexação a identificação, seleção e análise dos conceitos que deverão representar o conteúdo dos documentos, são partes precípuas de todo o processo.

Assim, os documentos selecionados para serem incluídos em uma base de dados deverão passar por um processo de análise de seus conteúdos, para que

possam ser representados de modo a operacionalizar a sua posterior recuperação.

De acordo com Gardin (1981) *apud* Alcaide *et al.* (2001) a análise documentária pode ser definida como um conjunto de procedimentos destinados a expressar o conteúdo dos documentos de forma que facilite a recuperação da informação. Para Ruiz (1992) *apud* Guimarães (2003) a análise documentária pode ser conceituada como: "conjunto de operações necessárias para a extração da informação contida nas fontes primárias de modo a prepará-la para sua posterior recuperação e utilização". Nestas duas definições a análise documentária deverá fornecer os subsídios operacionais indispensáveis para a representação do conteúdo dos documentos. A conseqüência imediata desta ação é que tais subsídios operacionais devem estar apoiados, segundo Cunha (1990) apud Alcaide *et al.* (2001), em duas fases:

- **Análise:** objetiva a identificação da estrutura do discurso do autor/produtor; e
- **Síntese:** objetiva atribuir ou mesmo extrair conceitos/descritores envolvidos na tradução do conteúdo do discurso analisado.

Para Guimarães (2003), a análise documentária também se divide em duas etapas:

- **Etapa analítica** - se subdivide em dois momentos: leitura técnica do documento visando identificar as partes com maior conteúdo temático e identificação de conceitos visando identificar as partes mais significativas tematicamente; e
- **Etapa sintética** – se subdivide em três momentos: seleção de conceitos onde os assuntos são postos em categorias (principais, secundários e periféricos), condensação documentária visando a elaboração de um resumo para o documento e representação documentária para traduzir o conteúdo temático do documento em linguagem de indexação.

Ainda discutindo a concepção proposta por Guimarães (2003), a análise documentária se desdobra em dois níveis:

- **Nível de análise formal:** análise dos aspectos extrínsecos aos documentos com a finalidade de identificação e localização; e
- **Nível de análise de conteúdo:** análise dos aspectos intrínsecos do documento, ou seja, representação temática do conteúdo dos documentos.

A análise documentária, que deve se desdobrar na representação do conteúdo dos documentos, tem na leitura do documento quer seja realizada por um indexador ou por um computador, um fator crítico de sucesso. Na percepção de Cintra (1983), "seja numa leitura parcial, seja numa leitura global do texto, o fato é que há um trabalho de identificação de descritores ou palavras-chave que representem o documento". A preocupação da autora está justamente na proficiência da leitura realizada pelo indexador, ela é de fato realizada com efetividade suficiente para cumprir o objetivo de representar o conteúdo do documento? Para Cintra existem dois problemas envolvidos nesta questão:

- **Representatividade efetiva das partes selecionadas em relação ao documento** – a autora acredita que nesta questão os centros de informação já o fazem sem maiores problemas, principalmente devido à imposição da rapidez; e
- **Processo de leitura realizado pelo indexador** – é operacionalizado com o objetivo de selecionar palavras-chave e descritores que representem o conteúdo de um dado documento compatível com a linguagem de indexação. Neste processo, dois procedimentos são usados para a atribuição de descritores: a apreensão instantânea e a apreensão por análise. Neste dois tipos de apreensão, o problema que se coloca é que a prática do indexador em uma determinada área do

conhecimento é crucial para tornar o processo de indexação mais rápido.

Somando-se a prática do indexador com o conhecimento das demandas e necessidades de informação dos usuários, a associação da experiência com a compreensão da demanda tornará mais fácil o trabalho de análise documentária, já que por apreensão instantânea ou por análise, as palavras-chave e descritores serão representativos para o processo de busca e recuperação da informação.

A representação do conteúdo dos documentos esbarra em uma tríade de termos: conhecer, informar e representar que Mari (1996) destaca: "só podemos representar um fato qualquer se dele temos algum conhecimento; igualmente, só podemos informar um fato se dele também temos algum conhecimento. Estas relações nos levam a admitir a existência de conhecer como uma condição necessária tanto para informar, como para representar".

Daí a relação estreita que deve existir entre a análise documentária e a representação do conteúdo dos documentos. Da mesma forma, torna-se indispensável a discussão sobre os caminhos para a depreensão dos descritores para a representação dos documentos. Sobre os termos conhecer, informar e representar, Mari (1996) acrescenta: "o que torna compreensível esta relação entre eles é o fato de podermos fazê-los equivaler a uma função correspondente, determinando critérios restritivos de sua utilização, com vistas a certos resultados a serem alcançados. Logo, ao afirmar que conhecer precede representar, estamos nos referindo à correlação entre duas ordens conceituais, onde um conjunto de resultados, a ser atingido com o ato de conhecer, venha a tornar-se necessário ao desempenho de representar".

Nesta outra situação, a representação do conteúdo dos documentos, fruto da análise documentária e com a finalidade de operacionalizar a recuperação da informação, se torna mais efetiva se houver um conhecimento prévio de como

poderão ser recuperados os documentos agora em processo de representação de seus conteúdos. Assim sendo, a associação da representação dos conteúdos dos documentos com o conhecimento por parte dos indexadores, dos requisitos informacionais dos usuários, influenciará positivamente a precisão da resposta na recuperação da informação.

Há também na literatura sobre o assunto, a possibilidade de otimização da etapa analítica da análise documentária, utilizando a Pragmática que é um ramo da Lingüística que se encarrega da linguagem que está além de seus aspectos sintáticos e semânticos interfrasais. Segundo Baranow (1983), na Pragmática a linguagem é analisada levando-se em conta o contexto extra-lingüístico da comunicação, ou seja, tão importante quanto o texto em si, é o conhecimento do contexto em que se insere o seu significado global. A possibilidade de utilização da Pragmática poderá facilitar a compreensão da relação praticamente hierárquica entre conhecer, informar e representar, além de otimizar o trabalho do indexador na análise e representação do conteúdo dos documentos.

Entre a análise documentária e a representação do conteúdo dos documentos, existe uma relação de causa e efeito que pode ser verificada pelo fato de que só é possível representar o conteúdo dos documentos a serem inseridos em uma base de dados, a partir de uma análise documentária que envolva a análise e a síntese do conteúdo destes. Entretanto, para que as duas fases da análise documentária possam ser efetivas na descrição do conteúdo, o emprego dos requisitos dos usuários do sistema será de fundamental importância para uma correta análise e posterior representação do conteúdo dos documentos a serem recuperados. É inegável que a análise documentária hoje é uma tarefa em que o indexador deve buscar, sistematicamente, traçar paralelos entre o documento e o usuário do sistema.

Entretanto, para que a análise documentária possa cumprir integralmente o seu objetivo de apoiar a representação dos conteúdos dos documentos, será

necessário considerar amplamente nas suas duas fases de análise e síntese, as necessidades de informação dos usuários do sistema. Desta forma, quando do processo de busca e recuperação da informação, os descritores atribuídos ou extraídos dos documentos serão mais fidedignos ao representar tematicamente o seu conteúdo, que será posteriormente recuperado em cumprimento aos requisitos dos usuários.

Finalmente, cabe citar a reflexão de Guimarães (2003): "cada vez mais se tem clara a necessidade de uma adequação entre metodologias de tratamento temático e as distintas realidades, em suas três dimensões fundamentais: a do documento, a do usuário e a da organização".

4.1.3- Linguagens de indexação

A linguagem pode ser definida como um sistema ou conjunto de sinais fonéticos dentre outros que se prestam a expressão do pensamento e do sentimento (Costa & Melo, 1989). Na definição proposta por Ferreira (1988), linguagem refere-se ao uso da palavra articulada ou escrita como meio de expressão e de comunicação entre as pessoas.

A definição oferecida por Cintra (1983) diz que "a linguagem é uma representação simbólica que expressa uma função psicossocial complexa" ou seja, é um "...sistema, uma organização relacional, onde cada elemento existe, na medida mesma em que se relaciona a outro ou a outros do mesmo conjunto".

Para o eminente lingüista americano Noam Chomsky, a linguagem é uma faculdade humana e funciona como uma verdadeira "propriedade da espécie" que varia pouco entre as pessoas. Segundo Chomsky (1998), os correlatos mais próximos da linguagem humana provavelmente encontram-se nos insetos. "O sistema de comunicação das abelhas, por exemplo, partilha com a linguagem humana a propriedade de 'referência deslocada', nossa habilidade de falar sobre

algo que esteja distante de nós no espaço e no tempo; as abelhas utilizam uma intricada 'dança' para comunicar a direção, distância e desiderabilidade de uma fonte distante de mel".

O autor comenta ainda que a faculdade da linguagem está presente em cada um dos aspectos da vida, do pensamento e da interação humanos e que ela é responsável pelo fato de os seres humanos sozinhos no universo biológico possuírem uma história, diversidade e evolução cultural complexa e rica e até mesmo sucesso biológico.

Dentre as inúmeras definições de linguagem, Vizcaya (2001) relata que não há como conceituá-la sem que se considere o seu papel, não só como expressão do pensamento, mas também da informação e mais especificamente das linguagens documentárias no âmbito do processamento da informação. Assim, a linguagem pode ser definida do ponto de vista de sua estrutura como um sistema de signos cuja estrutura possui dois níveis: organização e integração, Garvin (1963) *apud* Vizcaya (2001).

Neste momento cabe a apresentação da distinção entre linguagem natural e linguagem documentária. A primeira corresponde a uma linguagem que está no lado oposto da linguagem controlada, é a linguagem verbal, "...que embora seja um caso particular, constitui na verdade um sistema de signos de espectro tão amplo, que todos os outros sistemas de linguagem podem repassar de língua. Daí, porque, freqüentemente, o termo linguagem seja usado por língua, ou a expressão linguagem natural..." (Cintra, 1983). A segunda pode ser definida como instrumento intermediário ou de comutação que se realiza na tradução da síntese dos textos e das questões dos usuários (Cintra, 1994). A autora ainda acrescenta no seu trabalho de 1983 que a linguagem documentária decorre das dificuldades que a linguagem natural apresenta ao ser utilizada na descrição de documentos.

A definição de Slype (1991) para linguagem documentária diz que: "todo o sistema de signos que permite representar o conteúdo dos documentos com a finalidade de recuperar aqueles pertinentes em resposta as consultas sobre estes conteúdos. Assim sendo, a linguagem documentária não se refere a critérios empregados na busca documentária: autor do documento, língua do texto, ano de publicação, entre outros". O autor distingue dois tipos de linguagens documentárias:

- **Linguagens de indexação:** permitem representar o conteúdo dos documentos e das consultas de forma analítica; e
- **Linguagens de classificação:** empregadas geralmente para representar o conteúdo de forma sintética.

Segundo Rowley (2002), há dois tipos de linguagens controladas de indexação baseadas em assuntos:

- **Linguagens alfabéticas de indexação** – assim como nos tesauros e listas de cabeçalhos de assuntos, os termos que correspondem aos assuntos são os nomes dos assuntos, colocados em ordem alfabética. Exerce-se controle sobre os termos a serem usados e se indicam as relações entre eles, mas os próprios termos são palavras usuais; e
- **Sistemas de classificação** – cada assunto é representado por um código ou notação e possuem como finalidade principal localizar os assuntos em uma estrutura que mostra as relações que eles mantém entre si.

Tálamo; Lara & Kobashi (1992) afirmam que as linguagens documentárias são consideradas instrumentos de controle da terminologia de uma determinada área do conhecimento, atuando em dois níveis:

- **A)** Na representação da informação obtida pela análise e síntese de textos; e
- **B)** Na formulação de equações de busca da informação. Ao que acrescentam: "sendo assim, a questão fundamental que se encontra entre as linguagens documentárias e os documentos, diz respeito aos princípios que regulam a representação da informação".

As observações anteriores encontram ressonância nas observações de Vizcaya (1991), de que tradicionalmente o processamento da informação contribui para a aplicação das linguagens documentárias que são destinadas a expressar (traduzir) o conteúdo semântico dos documentos para uma linguagem artificial e especializada. Daí a factibilidade da proposta de Tálamo; Lara & Kobashi sobre o uso da terminologia para otimizar a construção de linguagens documentárias, em especial a linguagem específica do tesauro.

As linguagens documentárias são formadas por meio de termos que aparecem em certos recortes da literatura científico-tecnológica, onde são isolados do texto em que ocorrem a fim de que cada um deles possua definição própria, em uma tentativa de controlar a ambigüidade entre eles. Possuem ainda como função referenciar objetos, pessoas entre outros e as palavras assumem a função de etiquetas a serem coladas nas coisas a que se referem (Novellino, 1998).

No contexto das linguagens documentárias existe o que se chama de linguagens de indexação que são instrumentos padronizados na forma de vocabulário controlado (tesauro, cabeçalhos de assunto, lista de autoridades) que se destinam à indexação ou coleção de termos de indexação empregados em um sistema de informação. Como propõe Slype (1991), as linguagens de indexação permitem representar o conteúdo dos documentos e das consultas de forma analítica.

Segundo Rowley (1988), as linguagens de indexação são linguagens empregadas para descrever o assunto ou outros aspectos da informação ou de documentos em um índice e possuem três características básicas:

- **Controle da linguagem de indexação** – formado por uma categoria na qual são organizados cabeçalhos de assunto. Os termos são relacionados em uma lista; o indexador seleciona os termos da lista para indexar o documento; os termos selecionados vão formar os indicadores daqueles documentos (mais próximos do assunto abordado). A linguagem de indexação controlada é utilizada na indexação humana;
- **Linguagem de indexação livre** – todo termo que aparece no documento pode ser considerado como termo de indexação. Este tipo de indexação é mais usado na indexação automática; e
- **Linguagem de indexação natural** – emprega a própria linguagem do documento, podendo ser considerada um tipo de linguagem de indexação livre. Geralmente está baseada em termos apreendidos do título, resumo ou de outras partes do documento.

Para Slype (1991), a linguagem de indexação se divide em três tipos: linguagens livres, linguagens controladas e linguagens codificadas.

A linguagem livre é constituída sobre a base de indexação em linguagem natural de documentos registrados em uma coleção. Existem dois tipos de linguagem livre:

- **Lista de palavras-chave** – é constituída por uma coleção de palavras significativas ordenadas alfabeticamente. Não entram, desta forma, artigos, conjunções, pronomes, preposições, numerais, alguns verbos e advérbios. São extraídas de forma automática do título, do resumo e do texto completo dos documentos registrados em um dado sistema; e

- **Lista de descritores livres** – é constituída por uma coleção de conceitos apreendidos por meio de um processo intelectual, a partir dos documentos registrados em um dado sistema. Estes conceitos são expressos por palavras ou por expressões extraídas dos documentos.

A linguagem controlada é construída antes da indexação dos documentos de uma coleção. Existem dois tipos principais de linguagens controladas:

- **Lista de autoridades** – é constituída por uma coleção de conceitos destinados a representar, de maneira unívoca, o conteúdo dos documentos e das buscas em um dado sistema. Tais conceitos são expressos por palavras ou por expressões extraídas de uma lista finita estabelecida *a priori*. Somente os termos que figuram nesta lista podem ser empregados para indexar os documentos e as buscas; e
- **Tesauro** – lista estruturada de conceitos destinados a representar, de maneira unívoca, o conteúdo dos documentos e das buscas em um dado sistema, e a apoiar o usuário na indexação dos documentos. Os conceitos são extraídos de uma lista finita, estabelecida *a priori*. Somente os termos que figuram nesta lista podem ser empregados para indexar os documentos e as buscas. Possui uma estrutura semântica com relações entre os termos de equivalência, hierarquia e associação.

A linguagem codificada, assim como os sistemas de codificação em geral, utilizam tradicionalmente linguagem de classificação. Muitas linguagens controladas utilizam um sistema de notação denominado de topológico. A finalidade disto está em facilitar a localização dos descritores dentro de uma representação gráfica ou classificatória do tesauro, e jamais de substituir os descritores no momento da indexação dos documentos.

Wellisch (1995), se posiciona de maneira diferente sobre as linguagens de indexação. Segundo ele, todas as linguagens de indexação usadas para o

rearranjo da linguagem natural de um texto são na realidade linguagens controladas. Entretanto se submetem a um controle bem mais intenso do que a linguagem natural do homem, porque:

A) Na linguagem controlada pode-se utilizar apenas certas classes de palavras: substantivos, adjetivos, particípio e gerúndio, algumas preposições e conjunções, quase nenhum advérbio ou pronome e nenhuma interjeição;
B) Em sua sintaxe podem utilizar estas palavras somente em determinadas construções: não são empregadas frases completas com sujeito, verbo, objeto;
C) Em sua semântica, as palavras se restringem a seu significado, e se a palavra for ambígua deve haver a sua qualificação (com atribuição de sentido); e
D) Em seu pragmatismo, as palavras devem ser utilizadas somente em senso básico e primário e não em seu sentido metafórico, figurado ou simbólico.

Quem exerce este controle? A resposta dada por Wellisch (1995) recai no indexador, sobretudo quando segue as convenções e as regras da indexação, observando as limitações do uso de determinados termos, baseando-se em uma linguagem controlada chamada também de tesauro ou lista de cabeçalhos. Atualmente quando um texto é indexado, o que acontece é a aplicação maior ou menor do controle da linguagem de indexação.

A posição de Wellisch (1995) está associada com o fato de que a responsabilidade na busca da precisão de resposta em sistema de informação recai, em grande medida, na tarefa do indexador, sobretudo se as limitações do uso de determinados textos não estiverem respaldadas nas necessidades informacionais (requisitos) dos usuários do sistema.

No estudo de Rowley (2002) é apresentada uma tabela comparando as vantagens e desvantagens entre as linguagens de indexação controladas e não-controladas que é reproduzida a seguir:

Tabela 1 – Comparação entre as linguagens de indexação controladas e não-controladas

	Vantagens	***Desvantagens***
Linguagens de indexação não-controladas	Baixo custo; processo de buscas simplificado; possível fazer buscas no conteúdo total da base de dados; toda palavra tem valor de recuperação igual; sem erros humanos de indexação; sem demora na incorporação de novos termos	Maior carga de trabalho para o indexador; podem-se perder informações que estejam incluídas implícita mas não explicitamente no texto; ausência de vínculos do genérico para o específico; é preciso conhecer o vocabulário da disciplina
Vocabulário controlado	Resolve muitos problemas semânticos; permite que relações de gênero-espécie sejam identificadas; mapeia áreas do conhecimento	Custo alto; possíveis inadequações de cobertura; erro humano; possibilidade de vocabulário desatualizado; dificuldade de incorporar sistematicamente todas as relações relevantes entre os termos

Fonte Rowley (2002)

Esta comparação também pode ser encontrada em Lopes (2002).

4.1.4- Coerência e qualidade da indexação

A coerência da indexação diz respeito, segundo Lancaster (1998), à extensão do consenso quanto aos termos a serem empregados na atividade da indexação.

De modo semelhante, Slype (1991) diz que a coerência da indexação de um mesmo documento por dois indexadores é a proporção entre o número de descritores comuns e o número total de descritores comuns ou diferentes variando de 50 a 80%, segundo a qualidade do manual de indexação, a formação recebida pelos indexadores e a meticulosidade com que realizam a sua tarefa. Assim, a coerência pode ser medida da seguinte maneira:

- **Situação** - dois indexadores ou dois grupos de indexadores indexam o mesmo documento ou conjunto de documentos a partir de um mesmo tesauro e trabalhando de forma independente um do outro;

- **Procedimento** – conta-se separadamente para cada indexador ou grupo de indexadores: **I.** O número de descritores idênticos empregados pelos indexadores; **II.** O número total de descritores idênticos ou diferentes usados pelos dois indexadores ou grupos de indexadores; e **III.** A taxa de coerência é obtida por intermédio da proporção entre estes números.
- **Exemplo** – O indexador ou grupo de indexadores 1 apreendeu os descritores: A, B, C, D, E e F; O indexador ou grupo de indexadores 2 apreendeu os descritores: A, C, D, F, G e H. Foram apreendidos 4 descritores idênticos (A, C, D e F) de um total de 8 descritores (A, B, C, D, E, F, G e H). Desta forma, a taxa de coerência = 4/8 = 0,5 X 100 = 50% (Slype, 1991).

Como complemento à definição da coerência, há duas questões propostas por Lancaster (1998), a coerência interindexadores que significa a concordância entre indexadores e a coerência intra-indexador que diz respeito à coerência de um indexador para consigo mesmo.

Diante do conceito e das implicações decorrentes da consideração da concordância, a atribuição de determinados termos na atividade de indexação deve ter como requisito, o impacto da escolha na otimização da busca e recuperação da informação, sem que questões ligadas à ambigüidade ou mesmo imprecisão possam afetar todo o processo de indexação comprometendo a melhor resposta.

Citando um estudo realizado por Hooper (1965), Lancaster afirma que um alto nível de coerência interindexadores é muito difícil de se alcançar, já que a análise de 14 estudos diferentes permitiu a Hooper chegar a valores muito dispersos, algo entre 10 a 80% de coerência. Nos estudos em que foi possível recalcular a taxa de coerência os resultados variaram de 24 a 80%. Como

conseqüência disto, a pergunta que se coloca é: quais fatores influenciam a coerência da indexação?

Os fatores propostos por Lancaster são:

- **1. Quantidade de termos atribuídos** – considerando que os termos de indexação são atribuídos segundo um critério de prioridade, ou seja, levando-se em conta a representatividade dos termos face ao conteúdo do documento, nem sempre haverá uniformidade e consenso quanto à quantidade exata dos termos a serem atribuídos;

Esta situação poderia ser resolvida se critérios pudessem ser firmados com base nas expectativas dos usuários do sistema. Com isso, a concordância mudaria do foco exclusivo do indexador, para um compartilhamento com as necessidades de informação dos usuários, o que ocasionaria por conseqüência, a diminuição das discordâncias a partir de requisitos previamente fixados. Cabe ressaltar que a eliminação total das discordâncias é praticamente impossível, já que o processo de indexação possui, na sua essência, um caráter de subjetividade.

- **2. Vocabulário controlado versus indexação com termos livres** – não há dúvida que o emprego do vocabulário deve melhorar a coerência na representação do conteúdo temático dos documentos. Embora Lancaster (1993) ressalte que nem sempre esta relação seja direta, pois ao cotejar o vocabulário controlado com os textos dos documentos a serem indexados, o responsável pelo trabalho encontre dificuldades para identificar as semelhanças entre os termos controlados do vocabulário com os possíveis termos a serem atribuídos nos documentos. Ainda mais se existirem discrepâncias quanto à terminologia de determinadas áreas do conhecimento;

Deve-se considerar nesta questão que a coerência pode ser facilitada pelo uso de vocabulários controlados na indexação. Parece óbvio que a opção definitiva pelo uso de termos livres não parece ser uma saída confiável nos sistemas em que se procure estabelecer critérios para a promoção da coerência interindexadores. Contudo, o procedimento que parece ser mais racional está na utilização das duas possibilidades, ou seja, tanto o emprego do vocabulário controlado quando a situação o exigir, quanto o uso de termos livres a fim de buscar alternativas para a representação temática dos conteúdos dos documentos.

Este grau de flexibilidade proposto possibilita o ajuste de necessidades de informação (requisitos) coletados junto ao usuário, bem como a adaptação de termos que não estão previstos dentre os instrumentos da linguagem controlada.

- **3. Tamanho e especificidade do vocabulário** – vocabulários maiores serão mais específicos, o que por sua vez poderá dificultar a sua utilização coerente. Esta afirmação de Tinker (1966, 1968) apresentada por Lancaster (1993) é justificada pela especificidade encontrada em vocabulários muito grandes, já que as alternativas de significado por serem também mais numerosas, apresentarão obstáculos no alcance da coerência;

Nem sempre a especificidade levará à melhor resposta. A alternativa de solução que se apresenta, ainda está apoiada na adaptação do vocabulário controlado de um sistema em seus inúmeros instrumentos, às necessidades de uso da informação detectadas. Com isso, a adoção de requisitos previamente definidos, inclusive na construção e manutenção de vocabulários controlados, poderá apoiar a promoção da coerência em detrimento da grande especificidade.

- **4. Características do conteúdo temático e sua terminologia** – a suposição de que a maior coerência acontecerá a partir da indexação de

tópicos mais concretos, tais como nome de pessoas e objetos físicos, e que esta mesma coerência será reduzida quando da indexação de tópicos mais abstratos, é o que caracteriza a problemática em torno do fator 4;

- **5. fatores dependentes do indexador –** a coerência interindexadores está relacionada a uma série de questões intrínsecas de cada indexador envolvido no processo. Desta forma, discrepâncias acerca do conhecimento especializado de cada um deles, deverá ser um dificultador na promoção da coerência. Se os indexadores possuírem o mesmo nível de conhecimento especializado, certamente serão mais coerentes. O mesmo ocorre com indexadores mais experientes, por trabalharem a mais tempo com atribuição de termos e uso de instrumentos de auxílio no processo de indexação, serão, por conseguinte, mais coerentes do que indexadores iniciantes;

- **6. Instrumentos de auxílio para o indexador –** o compartilhamento dos mesmos instrumentos auxiliares no processo de indexação é um grande facilitador da coerência interindexadores. Não há como abdicar de instrumentos de apoio na tarefa da indexação. O uso de dicionários, glossários, manuais, etc., aumentam as possibilidades de atribuição de termos significativos na representação do conteúdo temático dos documentos;

Não só a utilização dos mesmos instrumentos de apoio na tarefa de indexação pelos indexadores vai ser suficiente para o alcance da coerência. A criação por consenso entre estes profissionais, de novos instrumentos em face de novas dificuldades a serem encontradas durante o trabalho, serão de importância crucial para a coerência interindexadores.

- **7. Extensão do item a ser indexado** – quanto menor a extensão do documento a ser indexado, menor será a quantidade de termos a serem atribuídos na sua representação. Então a menor quantidade de termos atribuídos concorrerá para que as discrepâncias na atribuição de termos interindexadores também seja reduzida. Citando Tell (1969), Lancaster (1993), concluiu que quando a indexação era realizada a partir do texto integral dos artigos, a coerência era menor quando se indexava a partir dos títulos e resumos destes artigos.

Todos os fatores apresentados podem ser cotejados na Figura a seguir:

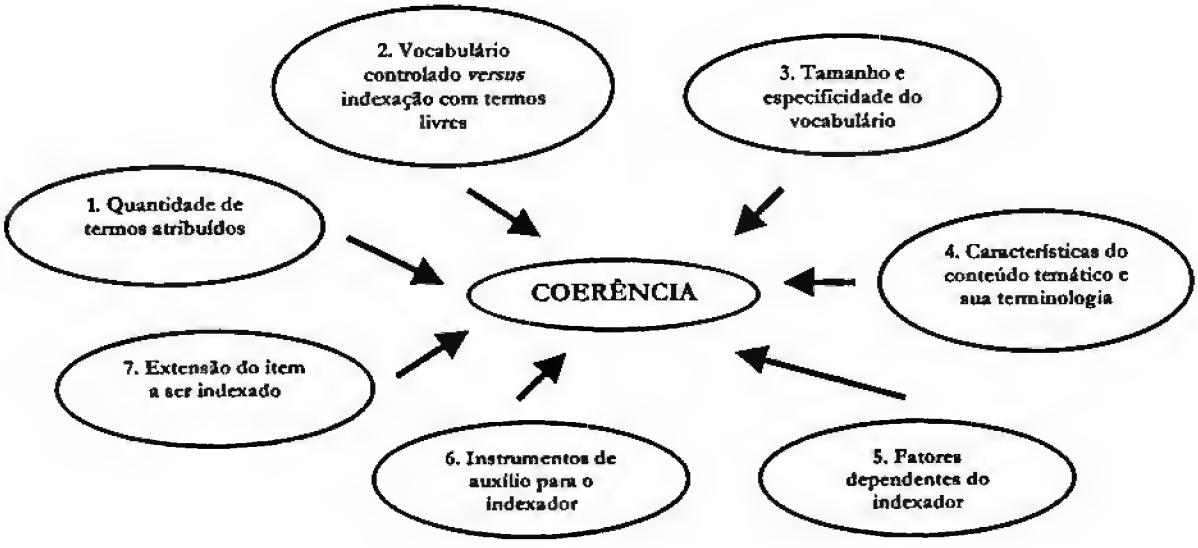


Figura 7 – Fatores que influenciam a coerência da indexação
(Fonte: baseado de Lancaster, 1998)

Sobre a qualidade da indexação, Slype (1991) relata que a exaustividade mede a qualidade na seleção dos conceitos de fatos significativos, ou seja, que contém informação pertinente para os usuários:

- Exaustividade muito reduzida ocasionará a não recuperação de documentos pertinentes;

- Exaustividade muita elevada ocasionará a recuperação de documentos que não contenham informação pertinente sobre os conceitos da consulta.

A exaustividade depende de:

- Política de indexação; e
- Qualidade do trabalho dos indexadores, em especial a sua capacidade de julgar o que é mais importante, ou seja, para detectar os conceitos implícitos.

A especificidade mede a qualidade na seleção dos descritores que correspondem aos conceitos inclusos no documento. Assim temos:

- Especificidade vertical: o descritor deve estar situado no mesmo nível de especificidade que o conceito;
- Especificidade horizontal: um conceito composto deve ser traduzido por um descritor pré-coordenado.

As especificidades horizontais e verticais dependem da:

- Riqueza do tesauro: só podem ser indexados os conceitos presentes na linguagem documentária; e
- Qualidade do trabalho dos indexadores e em particular a sua meticulosidade.

Segundo Lancaster (1998), a qualidade da indexação refere-se à consideração de diversos fatores ligados ao indexador, ao vocabulário, ao documento, ao processo e fatores ambientais, a coerência da indexação que influem de maneira geral naquilo que o autor chama de '*good' indexing* ou boa

indexação (aquela que possibilita a recuperação de itens de informação em uma base de dados, de fato úteis às demandas informacionais).

Cada um dos fatores que influenciam a qualidade da indexação estão dispostos conforme a Figura 8 apresenta:

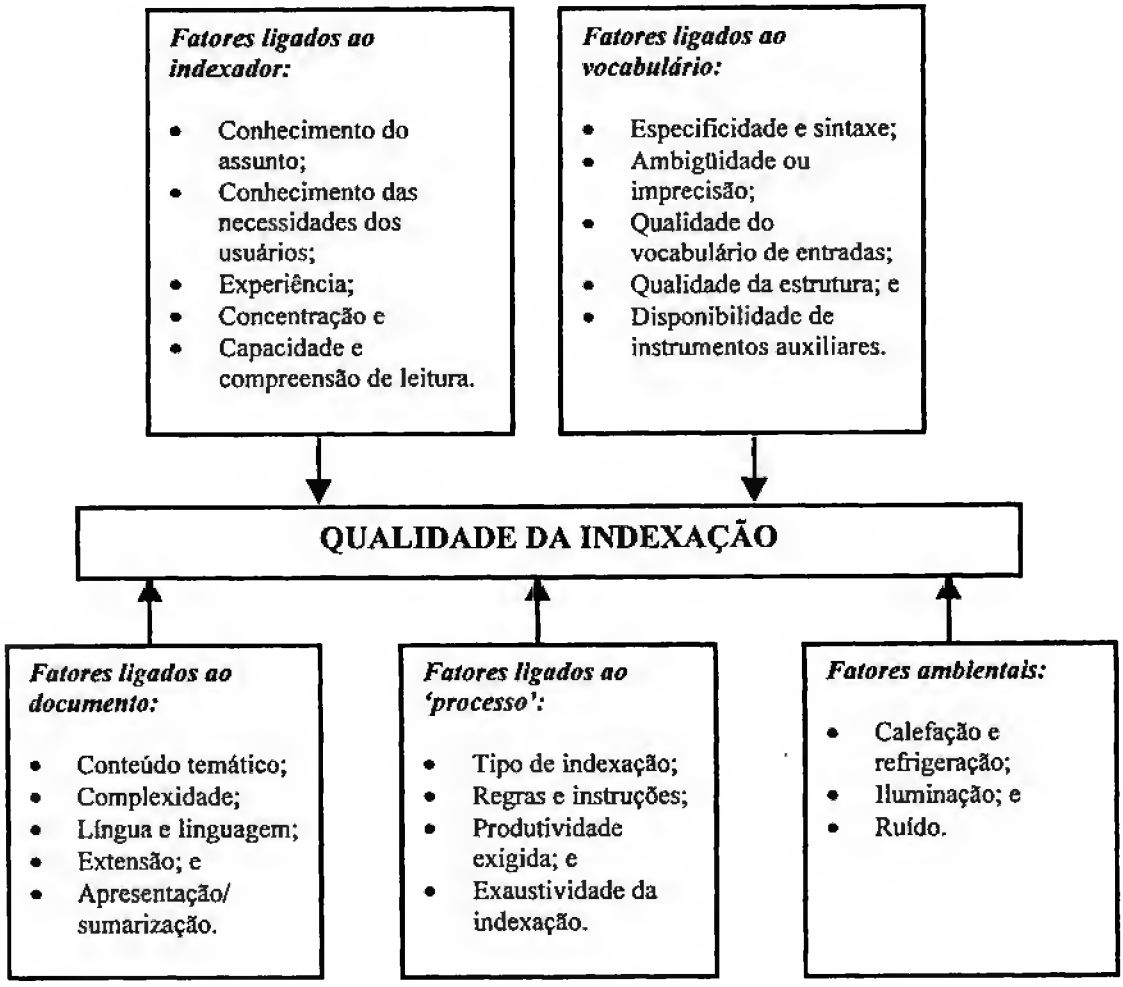


Figura 8 – Fatores que influenciam a qualidade da indexação
(Fonte: baseado de Lancaster, 1998)

4.1.5- Conclusão

O processo de indexação, por estar envolvido diretamente com a descrição e representação do conteúdo dos documentos, além de desempenhar um papel preponderante no processo de busca e recuperação da informação, tem na análise, descrição e representação dos conteúdos dos documentos seus fatores críticos de sucesso. Estas tarefas são capitais ao se definir a indexação, e, sobretudo, o seu papel no processo de busca e recuperação da informação.

A base de dados e o vocabulário do sistema (vocabulário controlado dos termos de indexação usados para disponibilizar o acervo documental) serão os elos de ligação entre a gestão do sistema e as necessidades informacionais dos usuários na criação da estratégia de busca.

A Figura 9 ilustra esta situação:

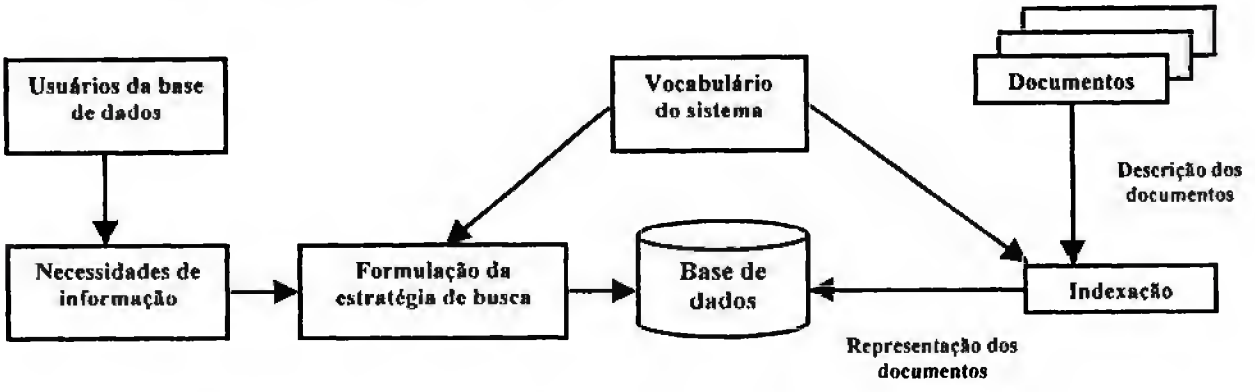


Figura 9 – A base de dados e as necessidades de informação
(Fonte: adaptado de Lancaster, 1998)

Sendo assim, a análise documentária passa a ser uma ferramenta indispensável ao processo de indexação e, para que ela possa cumprir integralmente o seu objetivo de apoiar a representação dos conteúdos dos documentos, será necessário considerar amplamente, nas suas duas fases de análise e síntese, as necessidades de informação dos usuários do sistema.

Como conseqüência, os indexadores atribuídos ou extraídos dos documentos serão mais fidedignos ao representar tematicamente o seu conteúdo, facilitando sua recuperação em cumprimento aos requisitos dos usuários.

Quanto à coerência da indexação, os fatores: quantidade de termos atribuídos, vocabulário controlado *versus* indexação com termos livres, tamanho e especificidade do vocabulário e instrumentos de auxílio para o indexador deverão ser geridos na perspectiva das necessidades informacionais do usuário, a fim de promover uma melhor resposta no processo de busca e recuperação da informação. No que concerne à qualidade da indexação, os fatores ligados ao indexador, ao vocabulário e ao documento deverão ser administrados com o foco nas necessidades de informação dos usuários, a fim de promover também uma melhor resposta, daí a importância do controle e avaliação destes dois elementos.

Enfim, a indexação definida por Cintra (1983) como a tradução de um documento em termos documentários para expressar o seu conteúdo deve ser operacionalizada por:

I. Leitura analítica do documento feita pelo indexador para identificar e selecionar indexadores que possam representar de forma fidedigna o seu conteúdo;

II. Controle e avaliação da coerência e qualidade da indexação; e

III. Consideração das necessidades de informação dos usuários.

A criação de valor para o processo de indexação deve passar pelo conhecimento proativo das necessidades dos usuários, o que deve proporcionar subsídios para a determinação dos requisitos a serem utilizados no âmbito do gerenciamento estratégico da informação.

4.2- A mineração de textos

O gerenciamento da informação tem se constituído em um grande desafio para as organizações, sobretudo frente a um mercado altamente competitivo que exige respostas rápidas às demandas do cliente. Com isso, a massa de dados que é encontrada nos sistemas de informação em grande parte das organizações não passa de um conjunto de dados que, de acordo com Amaral (2001), são inconsistentes, redundantes e pouco proveitosos para o processo de tomada de decisão. Encontrar padrões úteis nestes dados sem tratamento adequado e, por isso mesmo aparentemente sem valor, tem sido intitulado prospecção, descoberta de conhecimento em bancos de dados, mineração de dados, descoberta de conhecimento em textos, mineração de textos ou conforme as siglas em inglês *KDD* para *Knowledge Discovery in Databases* e *KDT* para *Knowledge Discovery in Text.*

No trabalho de Benoît (2002), a descoberta de conhecimento em base de dados é usada como sinônimo de extração da informação, mineração de conhecimento e mineração de dados. No trabalho de Delmater & Hancock (2001), a confusão persiste na definição dada para descoberta de conhecimento – primeiro componente da mineração de dados, usando sistematicamente instrumentos manuais e automáticos para detectar e caracterizar associações possível nos dados.

Com a mineração de textos e a descoberta de conhecimento em textos ocorre exatamente da mesma maneira que na mineração de dados, ou seja, os termos são usados como sinônimos na literatura. Este fato se deve, segundo Trybula (1999), pela incipiência dos estudos realizados. Para o autor, esta é uma boa oportunidade para os pesquisadores buscarem consenso quanto ao uso de definições e termos mais precisos, fato que ainda não ocorreu, conforme destaca Benoît (2002).

Por outro lado, é importante assinalar que a mineração é, na realidade, uma das técnicas de descoberta de conhecimento. Desta forma, a mineração de textos é parte do processo responsável pela aplicação de algoritmos de extração de padrões de dados (Wives & Loh, 2000). A Figura a seguir, resume o propósito da mineração de textos:

Figura 10 – Esquema da mineração de textos
(Fonte: adaptado de Tarapanoff *et al*, 2001)

Para Dixson (1997) *apud* Trybula (1999), o processo de mineração de textos pode ser conceituado como um meio de encontrar padrões interessantes ou úteis em um contexto de informações textuais não estruturadas, combinando alguma tecnologia de extração e de recuperação da informação, processo de linguagem natural e de sumarização/indexação de documentos. A chave para a mineração de textos está em obter um conhecimento anteriormente disperso em um grande volume de informações, sendo utilizado em computador para manipular documentos em repositórios eletrônicos de textos ou base de textos.

Segundo afirma Trybula (1999), as organizações vêm produzindo de 60 a 100% de informação a mais anualmente, o que deve levar a um aumento da ordem de 75% em investimentos em tecnologia da informação, principalmente no que se refere ao seu armazenamento. A partir do momento em que as bases de dados são formadas, faz-se necessário o desenvolvimento de mecanismos que permitam a mineração e a identificação de conhecimento "perdido" nos textos.

A mineração de textos possibilita o estabelecimento de ligações e compartilhamento do conhecimento entre as pessoas e as organizações. Com o uso de computadores com capacidade de processamento cada vez maior, aliado à necessidade dos governos e das corporações de gerir a vasta quantidade de informações, a mineração de textos é uma área que vem registrando um

crescimento muito grande nos últimos anos, justamente por se apresentar como alternativa na captura de padrões escondidos em grandes bases textuais. Sob outra ótica, a habilidade de mineração de bases textuais também vem se tornando cada vez mais importante à medida que a informação é produzida de forma eletrônica (há uma estimativa de que 80% do material escrito produzido, primeiramente é feito na forma eletrônica). A Figura 11 mostra os elementos envolvidos no processo de mineração de textos.

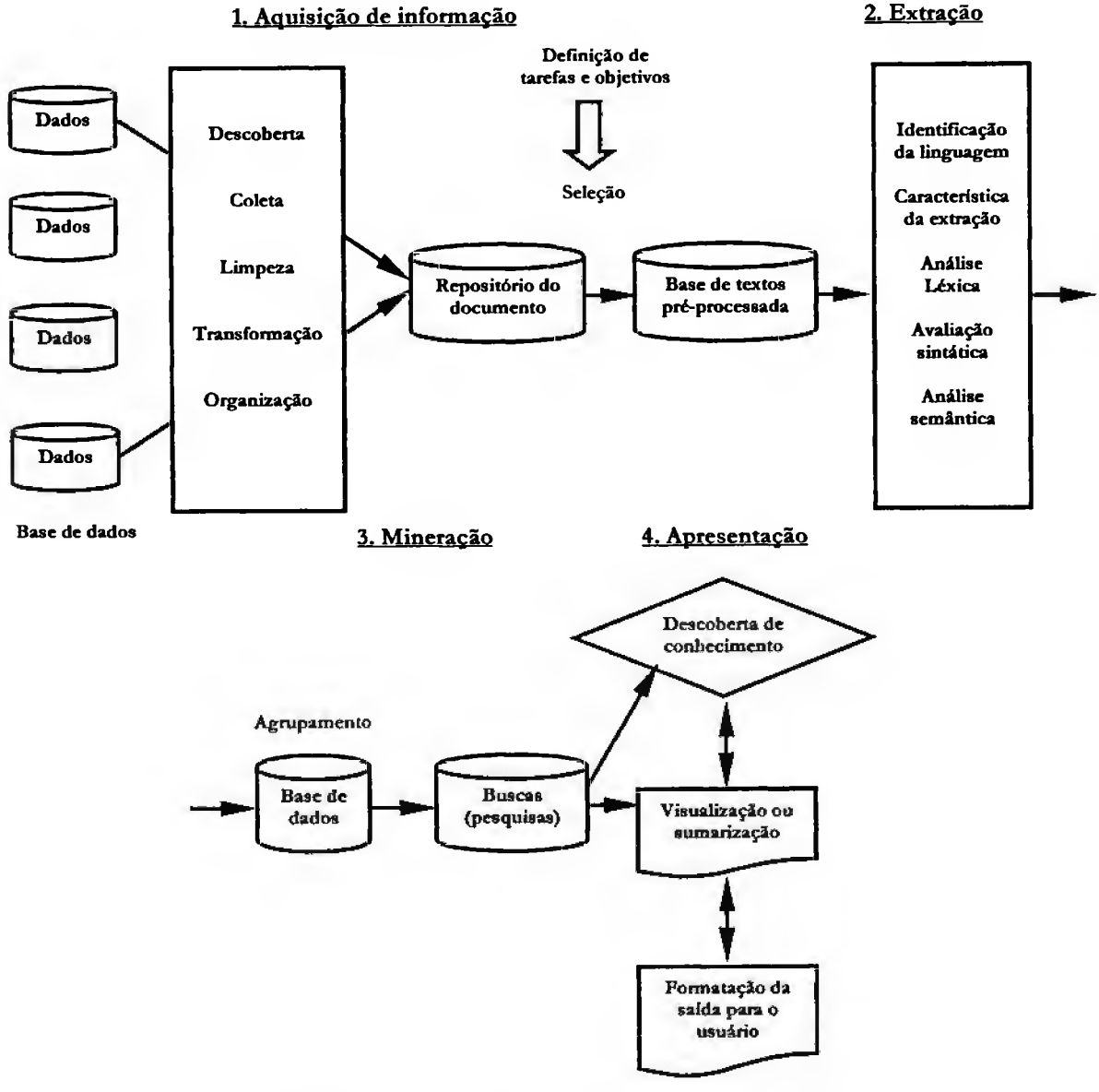


Figura 11 – Processo de mineração de textos
(Fonte: Trybula, 1999)

Para Feldman & Hirsh (1997) *apud* Wives (1999), a mineração de textos consiste na recuperação, filtragem, manipulação e resumo do conhecimento obtido de grandes fontes de informações textuais para apresentá-lo ao usuário por meio de gráficos, listas ou tabelas. Assim, o conteúdo de um conjunto de documentos poderá ser sumarizado, para gerar tabelas onde a relação semântica dos termos deverá ser apontada. Este resultado permite uma análise acurada do emprego e da factibilidade de cada descritor na representação dos conteúdos dos documentos, podendo ser empregado na potencialização de todo o processo de indexação.

Para Polanco & François (2000) a mineração de textos consiste na extração de informações sobre tendências ou padrões em grandes volumes de documentos textuais. Uma amostra significativa de informações é avaliada em textos contidos em bases de dados e em fontes de informação em linha. Os autores afirmam que a mineração de textos pode ser subdividida em cinco passos:

1 - Seleção de dados;
2 - Extração de termos e filtragem;
3 - Agrupamento de dados;
4 - Mapeamento dos agrupamentos ou visualização; e
5 - Resultado e interpretação.

A interpretação dos resultados por vezes permeia as demais fases, visto que em determinados casos a quantidade de informação é muito grande, exigindo como conseqüência, uma interpretação de modo a restringir o número de informações úteis (etapa de mineração).

O resultado do método de agrupamento, segundo Polanco & François (2000), pode ser empregado de duas formas:

I. Para sumarizar o conteúdo da base de dados considerando as características de cada agrupamento criado; e

II. Para apoiar outros métodos de avaliação de textos.

Assim, o conteúdo da base de dados poderá ser organizado na forma de agrupamentos, além de poderem ser representados graficamente. Com isso o mapeamento dos agrupamentos permitirá uma visão global da posição dos mesmos na indexação de textos e a sua interface com os demais agrupamentos.

As etapas básicas de um processo de mineração de textos são:

I. Definição de objetivos;
II. Seleção de um subconjunto de dados;
III. Pré-processamento ou limpeza dos dados, removendo ruídos e preparando os dados;
IV. Redução ou projeção dos dados (escolha de características relevantes para a análise);
V. Escolha da técnica, método ou tarefa de mineração;
VI. Mineração dos textos;
VII. Interpretação dos resultados, podendo, caso necessário, retornar aos passos anteriores do processo; e
VIII. Consolidação do conhecimento descoberto (documentação ou incorporação dos dados no sistema).

4.2.1- A mineração de textos e a mineração de dados

Todas as formas de mineração de dados objetivam descobrir padrões ainda desconhecidos nos dados. Esta técnica cada vez mais é apresentada como solução para a manipulação de gigantescas bases de dados e bases textuais.

A mineração de dados pode ser vista como o processo de descoberta de novas correlações, padrões e tendências significativas por meio da análise minuciosa de grandes conjuntos de dados estocados. O processo vale-se de tecnologias de reconhecimento, utilizando padrões/técnicas estatísticas e matemáticas.

Para Cabena (1997), a mineração de dados é a técnica de extrair informação previamente desconhecida e de máxima abrangência a partir de bases de dados, para usá-las na tomada de decisão. Aspectos específicos da mineração de dados incluem ainda, a investigação e criação de conhecimento, processos, algoritmos e mecanismos de recuperação de conhecimento potencial de estoques de dados (Norton, 1999).

Um determinado tipo de dado poderá ser processado com a finalidade de buscar padrões/tendências, dados textuais podem também ser processados com a mesma finalidade, ou seja, ser transformados em estruturas adequadas e serem processadas com métodos de extração do conhecimento (Amaral, 2001).

A mineração de dados tem sido utilizada para a identificação de informação em base de dados alfanuméricas. De acordo com Trybula (1999), estas bases de dados estão baseadas em campos relacionados ou estrutura de arquivos. Existe uma correlação entre os campos e as informações neles contidas, por exemplo o campo "nome" contém o nome das pessoas, o campo "endereço" contém seus endereços e assim por diante. A conexão de dados é uma propriedade fundamental das bases de dados. A correlação entre certos produtos e uma descrição dada por um determinado consumidor pode dar *insights* na decisão de compra. É por isso que, aliado ao crescente desenvolvimento da tecnologia da

informação, estão sendo desenvolvidas as aplicações e ferramentas da mineração de dados.

Já as bases textuais, segundo Trybula (1999), são coleções de documentos em linguagem natural, sem formato pré-definido para seus conteúdos como acontece com as bases de dados.

4.2.2- Tipologia da mineração de textos

Existem vários tipos de mineração de textos onde a descoberta e a extração de conhecimento podem se dar na forma de informações ou na forma de regras. Wives & Loh (2000) apresentam dezessete tipos, dos quais os quatorze principais estão dispostos a seguir:

- **Descoberta tradicional após extração:** os dados extraídos dos textos são formatados em base de dados estruturadas, onde com o apoio de técnicas de extração da informação e mineração de dados estruturados, serão descobertos dados úteis para o usuário;
- **Descoberta por extração de passagens:** utilizando regras gerais ou mesmo criando as suas próprias regras, o usuário poderá selecionar uma passagem de um texto que contenha, por exemplo, o seu objetivo. Neste tipo de descoberta, detalhes de informação podem ser obtidos sem que o usuário tenha que ler o texto na íntegra, mas as passagens selecionadas devem ser lidas e interpretadas;

- **Descoberta por análise lingüística:** este tipo de descoberta se dá por meio de análises lingüísticas a nível léxico, morfológico, sintático e semântico. É possível descobrir generalizações escondidas por intermédio da análise de padrões sintáticos. Como exemplo, os autores tomam: ocorrendo num texto as frases 'x compra y' e 'z compra y' e sabendo-se que numa hierarquia auxiliar de conceitos está definido que w é "pai" de x e z, então se deduz que 'w compra y'.
- **Descoberta por análise de conteúdo:** propõe a investigação lingüística dos textos para apresentar informações sobre o seu conteúdo aos usuários. O conteúdo pode ser o tema ou assunto do texto, ou até mesmo um índice ou resumo;

- **Descoberta por sumarização:** utiliza-se da abstração das partes mais importantes do texto com ênfase no resumo ou no sumário. Este tipo de descoberta possui várias abordagens, dentre elas a de Mike *et al.* (1994) *apud* Wives & Loh (2000). Consegue gerar resumos em tempo de execução por meio de interações com os usuários, sendo que o tamanho do resumo, bem como as suas partes, podem ser definidos pelo usuário;

- **Descoberta por associação entre passagens:** objetiva encontrar automaticamente em um ou vários textos correlações entre conhecimento e informação. É aplicado na definição automática de *links* nos sistemas de hipertextos, apresentando aos usuários partes do texto que tratam do mesmo assunto;
- **Descoberta por listas de conceitos-chave:** consiste na apresentação de uma lista contendo os conceitos principais de um dado texto. Neste tipo de descoberta, o significado do texto é determinado por uma análise de palavras-chave mais importantes, ao invés de uma simples leitura linear (Moscarola, 1998 *apud* Wives & Loh, 2000). Na identificação das palavras-chave podem ser usadas técnicas simples de extração de termos mais freqüentes;

- **Descoberta de estruturas de textos:** por meio da estrutura de um texto é possível se chegar ao seu significado, já que ele é uma unidade coesa com frases que lhe dão significado. A coesão é obtida com referências, conjunções e relações semânticas (Morris & Hirst, 1991 *apud* Wives & Loh, 2000). Neste tipo de descoberta, as coesões léxicas de uma dado texto são analisadas e o resultado são cadeias de termos relacionados que contribuem para a continuidade do significado léxico;
- **Descoberta por agrupamento ou generalização:** consiste na separação automática de elementos em classes identificadas durante o processo. A interpretação das classes é feita pelos usuários e é usada para avaliar a associação entre os termos, a estrutura das coleções por análise terminológica e a informação qualitativa sobre as diferenças e semelhanças entre os componentes de cada classe;
- **Descoberta por descrição de classes de textos:** por meio de um agrupamento de documentos textuais um tema, ou mesmo assunto associado ao agrupamento, a descoberta por descrição deverá facilitar a descoberta das características principais desta classe, a fim de que possa identificá-la para os usuários e diferenciá-la das outras. Este tipo de descoberta se diferencia da descoberta por listas de conceitos-chave, por descobrir características comuns em vários agrupamentos de textos e não em apenas um texto;
- **Descoberta por recuperação de informações:** a recuperação da informação permite aos usuários encontrar soluções por analogias. Assim sendo, a recuperação da informação auxiliará apresentando documentos com visão geral dos assuntos ou mesmo parte de documentos com detalhes de informação. Por filtragem, a recuperação da informação contribui 'garimpando' documentos de interesse dos usuários, sem que este formule consultas;
- **Descoberta por associação entre textos:** procura correlacionar as descobertas presentes em textos diferentes, por meio do seu conteúdo e significado. Na associação entre textos, a interpretação semântica é

indispensável. Segundo Davies (1989) *apud* Wives & Loh (2000), existe muita informação publicada e conhecida, mas algumas conclusões a partir destas informações só poderão ser descobertas ao se recuperar estes documentos e estabelecer as suas conexões lógicas;

- **Descoberta por associação entre características:** relaciona tipos de atributos em textos com o uso de técnicas de correlação estatística diretamente sobre partes do texto. Ferramentas de descoberta procuram encontrar padrões na coleção de documentos por análise de distribuições de palavras-chave. Os padrões interessantes geralmente são subconjuntos cujas distribuições são diferentes do conjunto todo; e
- **Descoberta por hipertextos:** a descoberta é exploratória, além de ser realizada por meio de mecanismos de navegação (*browsing*). As técnicas de recuperação da informação atualmente estão mais voltadas para o processo de recuperação do que para a compreensão. Neste sentido, os sistemas de hipertexto podem facilitar as novas descobertas, permitindo ao usuário complementar seu conhecimento com informações adicionais.

4.2.3- Conclusão

A mineração de textos, por permitir a análise de uma amostra significativa de informações contidas em grandes base textuais e em fontes de informação em linha, é extremamente útil na descoberta de padrões inesperados nos dados. O gerenciamento de grandes massas de dados tem trazido uma série de transtornos para as organizações. Não se tem notícia, utilizando-se métodos tradicionais de análise de dados, de organizações que tenham obtido resultados satisfatórios na transformação de dados redundantes em informação com valor agregado útil ao processo decisório.

Por isso, as ferramentas de mineração de texto são cada vez mais utilizadas para resolver os problemas advindos da grande quantidade de textos

que se acumulam nas bases de dados e na memória dos computadores das instituições mundo afora.

Em termos práticos, a mineração de textos no âmbito do processo de busca e recuperação da informação, poderá ser pensada como uma ferramenta a ser empregada na busca da melhoria das respostas nestes sistemas. A avaliação desta possibilidade poderá ser materializada a partir da utilização do índice de precisão para aferir a performance da ferramenta nesta tarefa, bem como compará-la aos instrumentos tradicionais, utilizados hoje.

Outra possibilidade que pode ser descortinada, diz respeito à utilização da mineração de textos na extração automática de termos dos documentos das bases textuais, a fim de compor listas de palavras significativas validadas pelos indexadores que, por sua vez, poderão ser úteis para enriquecer os instrumentos de apoio ao processo de indexação, tais como o tesauro e a lista de palavras-chave usada em um sistema de recuperação da informação.

Por fim, a mineração de textos apenas fará sentido se aplicada a uma situação concreta, onde seja possível sua utilização para o aprimoramento e busca de valor adicional em um dado processo. A situação concreta está consubstanciada na verificação do ganho de precisão no processo de busca e recuperação da informação que a utilização da mineração de textos poderá trazer.

4.3- O processo de busca e recuperação da informação

O processo de busca e recuperação da informação pode ser conceituado como o processo de localizar documentos e itens de informação que tenham sido objeto de armazenamento, com a finalidade de permitir o acesso dos usuários aos itens de informação, objetos de uma solicitação. A recuperação da informação se dá pela comparação do que se solicitou com o que está armazenado, bem como

com o conjunto de procedimentos que este processo envolve (Belkin & Croft 1987).

A recuperação da informação possui limitações associadas à necessidade de informação, entendida como elemento-chave para a compreensão do motivo pelo qual os usuários se envolvem com o processo de busca e recuperação da informação (Le Coadic, 1994). As necessidades de informação acabam por gerar determinados graus de imprecisão, ou seja, incapacidade de um sistema de informação de recuperar documentos úteis frente à solicitação do usuário, sobretudo se há negociação com o usuário (Foskett, 1996). Sobre esta questão particular na busca e recuperação da informação, os requisitos do processo podem ser definidos pelo lado do usuário como motivação, que culmina na expressão de sua necessidade informacional. Na outra ponta do processo, a recuperação daquilo que foi demandado deverá se aproximar, o máximo possível, desta expectativa ou demanda informacional.

Todavia, a imprecisão acaba por se manifestar desde a entrevista de referência, onde nem sempre a necessidade de informação expressa pelo usuário é totalmente precisa. Segundo Grogan (1995), "os usuários prováveis que julgam que, para lidar com o problema que lhes diz respeito, precisam conhecer alguma coisa, avançaram para a segunda etapa da caminhada rumo à solução. Nesse ponto, talvez a sua necessidade de informação seja vaga e imprecisa, ainda que não necessariamente. Provavelmente, porém, ainda não estará nem formada e certamente nem expressa".

Estas restrições que envolvem o processo de busca e recuperação da informação acabam por influenciar de maneira decisiva todos os aspectos relacionados às características de um sistema de recuperação da informação, incidindo diretamente na utilidade da informação a ser recuperada, conforme pode ser visualizado na Figura 12.

De toda a forma, o monitoramento sistemático das necessidades de informação dos usuários poderá abrir caminhos para minimizar as restrições que envolvem o usuário (requisitos), o mecanismo de busca e o conjunto de informações recuperadas em consonância com os requisitos pré-determinados.

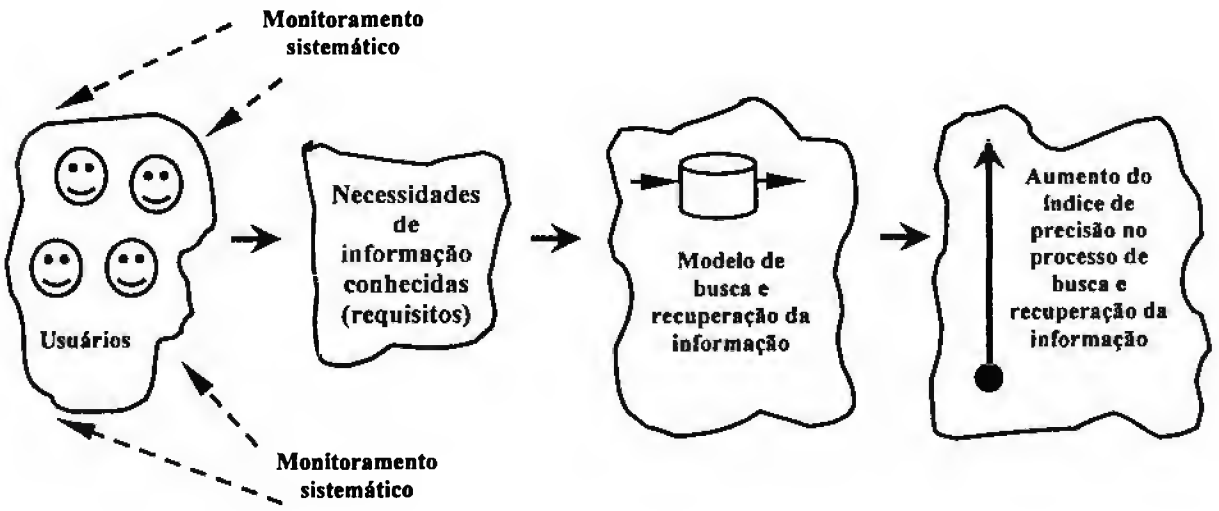


Figura 12 – A influência do monitoramento das necessidades de informação dos usuários na busca e recuperação da informação

Em um processo de referência, fica clara a posição da recuperação da informação como elemento vital. A fim de possibilitar uma compreensão mais adequada deste processo serão apresentados de modo esquemático os oito passos propostos por Grogan (1995), para a seqüência lógica das etapas decisórias que consubstanciam o 'processo normal de referência'

Etapa 1 – Problema: etapa que desencadeia simultaneamente o processo de referência e o processo de busca e recuperação da informação. Elemento gerador que necessita ser expresso e compreendido para que a busca de soluções seja factível. "A fonte do problema pode ser interna ou externa. Um problema externo

decorre do contexto social ou pelo menos situacional do indivíduo; um problema interno é de origem psicológica ou cognitiva, surgindo na mente da pessoa";

Etapa 2 – Necessidade de informação: elemento que necessita ser amplamente considerado para que a busca da solução para o problema possa se concretizar. Para que a necessidade de informação passe a integrar uma etapa decisória será necessária a sua expressão. A partir daí, o profissional da informação poderá tomar conhecimento do que o usuários necessita e predeterminar requisitos à satisfação da demanda. Isto impactará em grande parte na efetividade do processo de busca e recuperação da informação. "Há naturalmente várias maneiras de descobrir o que se deseja: observação, ensaio e erro, experimento; perguntar a alguém; procurar por si mesmo. O usuário potencial que experimenta uma das três primeiras opções e consegue ser bem-sucedido deixa de ser usuário potencial";

Etapa 3 – Questão inicial: neste momento a necessidade de informação será expressa por meio de uma construção lógica, moldada a partir da linguagem, aonde a demanda é finalmente expressa. Esta etapa é também decisiva para o processo de busca e recuperação da informação. No pensamento de Diaz (1986), não é possível que A e B – no nosso caso o usuário e o profissional da informação – percebam a realidade da mesma forma, já que a dinâmica mental interna de A e de B se vale de repertórios diferenciados de conhecimento, experiências, valores, crenças e atitudes, além de diferenças de signos que influenciam a percepção. Diaz (1986), ressalta ainda que A e B, não obstante, contem com habilidades perceptivas diferentes, um deles ouve melhor do que o outro ou mesmo B enxerga melhor do que A, etc (Figura 13). Desta forma, não há como refutar a influência da dinâmica da comunicação interferindo decisivamente no processo de busca e recuperação da informação;

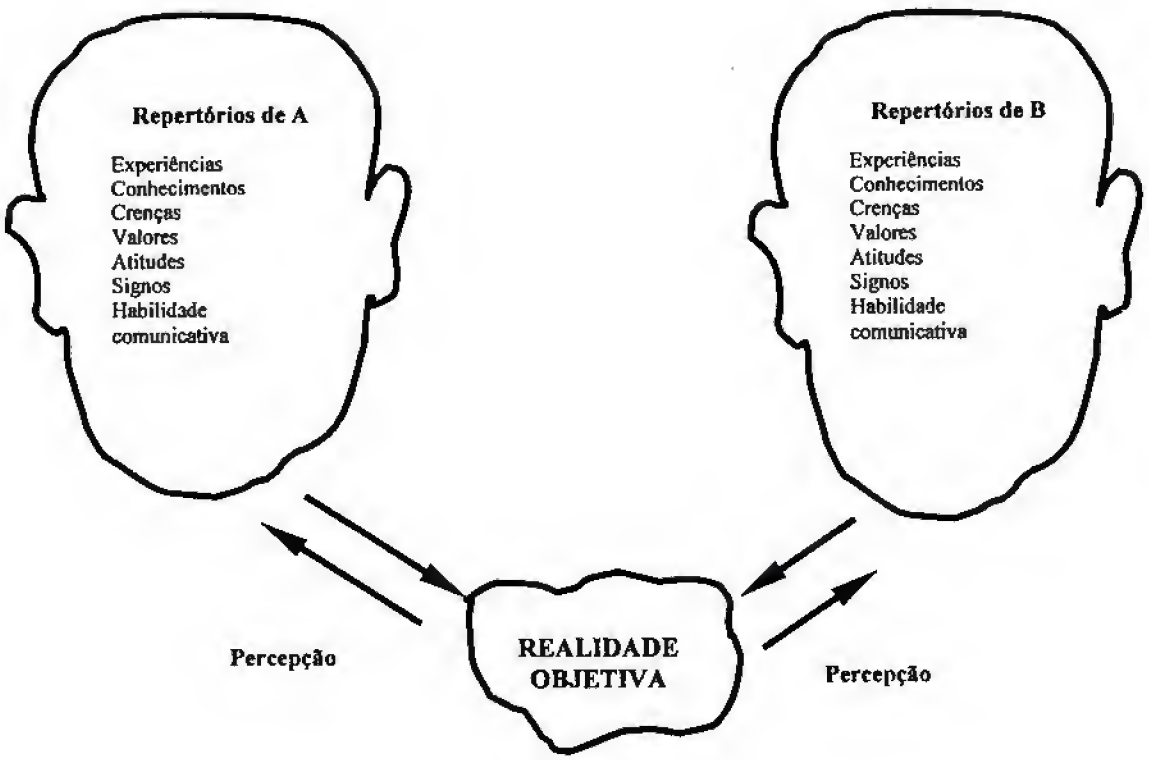


Figura 13 – Repertórios de A e de B
(Fonte: Diaz, 1986)

Etapa 4 – Questão negociada: representa o ajuste necessário à compreensão clara da demanda informacional. Estarão envolvidas aqui, peças importantes do processo de comunicação humana. De acordo com Diaz (1986), a troca de mensagens entre A e B conjuntamente com os processos de percepção, decodificação e interpretação, acabam por formar novos significados, já em parte compartilhados. Estes novos significados interagem de forma dinâmica com os significados iniciais, construídos na formulação da questão inicial, modificando-se por meio de diversos fatores.

A Figura 14 da página 74 apresenta esquematicamente esta questão.

Etapa 5 – Estratégia de busca: vai englobar decisões a cerca de como a questão inicial formulada e negociada (etapas 3 e 4) vai ser trazida ao acervo de informações. Isto significa dizer como se efetuará a busca e a recuperação em

uma base de dados, por exemplo. Neste instante, a decisão deverá estar de acordo com a complexidade da questão, bem como com as características da base, de modo a trazer na recuperação, informações relevantes para o consulente (usuário). De um modo mais amplo, as decisões a serem tomadas na busca e recuperação da informação deverão apontar caminhos possíveis para o êxito na satisfação da demanda expressa, por meio de requisitos definidos na formulação e negociação da questão. Para Grogan (1995), duas decisões são preponderantes: a análise minuciosa do tema objeto da questão, considerando suas relações e conceitos, transportando-os para uma linguagem de acesso ao acervo informacional. A segunda decisão refere-se ao desejável conhecimento aprofundado das várias fontes de informação disponíveis;

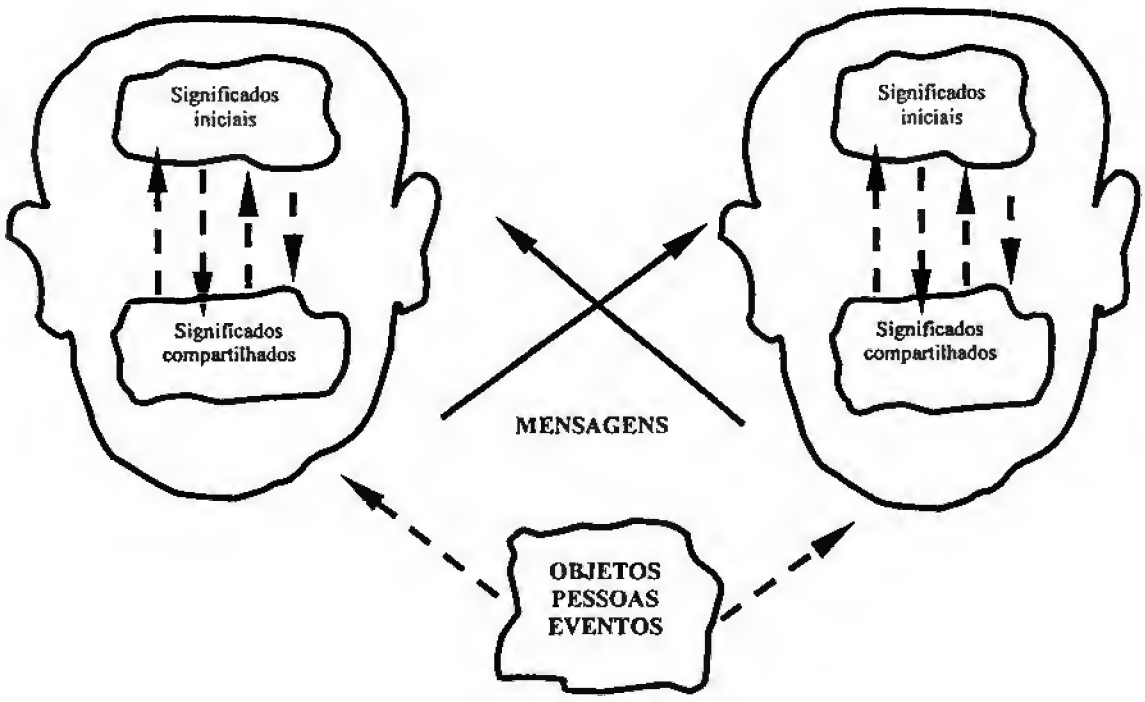


Figura 14 – Significados iniciais e compartilhados
(Fonte: Diaz, 1986)

Etapa 6 – Processo de busca: refere-se à concretização da busca no acervo de informação. Mais uma vez Grogan (1995) chama atenção para o fato de

que as buscas devem possuir uma estratégia flexível o bastante para mudanças necessárias durante o curso da ação, desta maneira será possível no processo de busca, aproximar-se dos requisitos definidos na formulação e negociação da questão;

Etapa 7 – Resposta: teoricamente a resposta como resultado final, poderia ser considerada como a última etapa do processo. Entretanto, a resposta acaba se constituindo como base para novos ajustes entre a demanda expressa em requisitos e o sucesso no atendimento destas expectativas. Daí considerarmos, conforme Grogan propõe, a resposta infrutífera como resposta legítima que poderá ensejar redefinições e mudanças de estratégias e decisões, inclusive com reavaliações nos recursos disponíveis para representação temática do conteúdo das informações disponíveis no acervo. Esta retroalimentação é própria dos sistemas e necessária no aperfeiçoamento do processo de busca e recuperação da informação; e

Etapa 8 – Solução: também a solução está condicionada à retroalimentação do processo de busca e recuperação da informação, realimentação que tem como alvo o reforço no atendimento dos requisitos pré-determinados.

De acordo com Caro; Cedeira & Travieso (2003), o atual interesse pela investigação sobre a interação entre usuário e sistema de busca e recuperação da informação, fomenta uma aproximação entre um conjunto cada vez maior de usuários interessados em mecanismos de recuperação que se adaptem às suas necessidades, ou seja, acesso a ferramentas interativas e amigáveis que proporcionem acesso à informação multimídia e favoreçam diferentes formas de busca.

Na construção de tais sistemas é condição precípua o conhecimento da demanda dos usuários e como se comportam durante a busca no uso das redes

de informação. Paralelamente, já há uma mudança no conceito da natureza da recuperação da informação. Não existe estratégia de busca a não ser a partir das necessidades de informação dos usuários (com estado anômalo de conhecimento). Da mesma forma, os elementos de saída de um sistema de busca e recuperação da informação, não são mais documentos potencialmente relevantes, mas o julgamento da informação por parte dos usuários cujo estado de conhecimento tenha se modificado durante a interação. Este posicionamento vem reforçar a argumentação proposta por Grogan (1995), acerca das duas últimas etapas decisórias que consubstanciam o 'processo normal de referência'.

Na perspectiva tradicional, o objetivo central da investigação no âmbito da busca e recuperação da informação está na melhoria contínua das técnicas de recuperação e nos métodos de representação da informação, de forma que se facilite a equiparação entre buscas e documentos. Caro; Cedeira & Travieso (2003), acrescentam também que os trabalhos de pesquisa neste âmbito simulam um contexto de busca artificial, controlada, formada por quatro componentes:

- Um conjunto de documentos;
- Um sistema de armazenamento e recuperação da informação;
- Um conjunto de temas que se concretizam em enunciados de busca; e
- Um conjunto de julgamentos sobre a utilidade dos documentos para estes enunciados (Caro; Cedeira & Travieso 2003).

Uma das tarefas fundamentais de um modelo de gerenciamento da informação deve incluir, na concepção de McGee & Prusak (1994), a determinação de como os usuários poderão acessar informações necessárias e o melhor lugar para seu armazenamento. Esta concepção deverá influir diretamente no modelo de uso da informação adaptado de Koblas (1995), representado na Figura 15:

Figura 15 – Modelo de uso da informação
(Fonte: adaptado de Koblas, 1995)

No processo de busca e recuperação da informação, não há como por comparação entre o que foi solicitado (após a formulação da questão e durante a sua negociação) e o que está armazenado, como proposto por Belkin & Croft (1987), que seja possível concluir o processo de recuperação da informação, sem que requisitos sejam estabelecidos com base nas demandas expressas pelos usuários. Estes elementos são vitais para que um modelo de busca e recuperação da informação obtenha sucesso.

4.3.1- Sistemas de recuperação da informação

A recuperação da informação é reconhecida como a recuperação de referências de documentos em resposta às solicitações (demandas expressas por informação). Já os sistemas de recuperação da informação dizem respeito a um sistema de operações interligadas para identificar, dentre um grande conjunto de informações (uma base de dados, por exemplo), aquelas que são de fato úteis, ou seja, que estão de acordo com a demanda expressa pelo usuário.

Segundo Rowley (2002), os sistemas de recuperação da informação quase foram usados como sinônimo para computadores, mas os sistemas baseados em papel, como os de fichas e arquivos de documentos existem e já estavam em voga antes do advento da informática e dos computadores. Para a autora os sistemas de recuperação da informação são compostos por três etapas: a

indexação; o armazenamento; e a recuperação, além de poderem ser divididos em cinco tipos diferentes.

A Tabela 2 apresenta os diferentes tipos de sistemas de recuperação da informação:

Tabela 2 – Diferentes tipos de sistemas de recuperação da informação

	Características dos usuários	*Ambiente*	*Tarefas*	*Tecnologia*
Serviços de busca em linha	Usuários especialistas e gerentes de informação	Escritório, biblioteca universitária, centro de informação empresarial	Recuperação de informações, importação de informações e integração com outros documentos	Variedade de estações de trabalho; configurações antigas com ligação direta com o serviço; mais aplicações na ponta; vínculos na Internet
Cederrom	Depende da base de dados – pode incluir crianças, usuários em geral de bibliotecas públicas, usuários especializados e outros	Biblioteca, aeroporto, residência, escritório	Recuperação de informações, importação de informações e integração com outros documentos	Freqüentemente multimídia, interface gráfica, mouse
Internet	Internautas – predomínio de docentes, estudantes e pessoas do sexo masculino	Lugar de estudo/trabalho, residência	Comunicação por correio eletrônico, vendas, transferência de arquivos	Microcomputadores de mesa e portáteis, com teclado, monitor e mouse
Catálogos em linha de acesso público	Usuários de biblioteca – o perfil depende do tipo de biblioteca	Biblioteca, escritório/residência , outros locais públicos	Estritamente definidas – identificação da disponibilidade de livros, busca de informação	Às vezes telas grandes, telas sensíveis ao toque, teclados especiais, mas também com acesso por meio de equipamento comum de escritório; acesso remoto e local; podem usar estações de trabalho diferentes
Sistemas de gerenciamento de documentos	Usuários institucionais com alguma experiência em comum do sistema, e objetivos e tarefas comuns	De escritório, mas pode também estender-se para operação móvel e uso em unidades de produção em, p.ex., trens e carros	Consulta a arquivo da empresa na execução de tarefas ligadas ao trabalho	Estações de trabalho ligadas a uma poderosa central de processamento; algumas aplicações serão de ponta

Fonte: Rowley (2002)

Na concepção apresentada por Robertson (1981), os sistemas de recuperação da informação podem ser entendidos como um conjunto de regras e procedimentos executados a partir da ação humana e/ou máquinas, englobando as seguintes atividades:

- **Indexação** (construção da representação do conteúdo dos documentos);
- **Formulação da busca** (formulação da questão que deve representar as necessidades de informação);
- **Busca** (confrontação das representações dos conteúdos dos documentos com a questão formulada para representar as necessidades de informação);
- **Retroalimentação ou *Feedback*** (repetição de uma ou mais operações ou modificações introduzidas nas respostas, a fim de avaliar os resultados de alguns dos processos relacionados à recuperação da informação); e
- **Construção da linguagem de indexação** (geração de regras de representação do conteúdo dos documentos).

No âmbito da definição de sistemas de recuperação da informação, Robertson (1981) apresenta ainda os conceitos de documento, demanda/questão e usuário:

Documento: é, em tese, considerado como sinônimo de texto em lingüística, isto é, ele deve descrever qualquer "parte" da lingüística de forma que o mesmo possa ser considerado como uma unidade, fato que comprova a utilização nos testes de efetividade de sistemas de recuperação da informação, as produções científicas como unidades, sobretudo nos testes empregados na década de oitenta;

A demanda (questão): é comumente utilizada para dar significado a uma necessidade de informação, entretanto com o desenvolvimento de sistemas de recuperação em linha, adquiriu significado geral de pesquisar/demandar. Robertson (1981), acrescenta que o ato de demandar é estimulado por uma necessidade especifica de informação e que o processo não é linear, uma vez que

a percepção do usuário quanto à sua demanda pode mudar dependendo da sua interação com o sistema empregado para a recuperação da informação; e

Usuário e demandante são sinônimos: ao contrário disto, existe uma distinção clara para Robertson entre um sistema que permite a recuperação de conhecimento recente ou em um sistema de disseminação seletiva da informação (DSI) e aqueles que permitem a recuperação de informações retrospectivas. Em termos do mecanismo empregado por estes sistemas, no caso da recuperação retrospectiva, a demanda é feita considerando a necessidade de identificação de um documento que não se enquadre na coleção de documentos correntes[10]. Já em um sistema de disseminação seletiva da informação, repetidas pesquisas são feitas no acervo de documentos, tendo em vista as adições feitas e o período de tempo em que as mesmas ocorreram, a importância maior será dada aos documentos mais recentes.

Enfim, qual é a proposta de um sistema de recuperação da informação? O que ele pode de fato realizar? Robertson (1981) propõe fornecer respostas para cada uma das demandas, mesmo sabendo que esta não é uma tarefa simples, já que cada demanda tem suas próprias peculiaridades.

Partindo do princípio de que o processo de recuperação da informação pretende satisfazer uma necessidade de informação, podemos descrever como função dos sistemas de recuperação da informação a seguinte premissa: "levar ao usuário/demandante o documento certo que irá satisfazer a sua necessidade específica de informação". Assim sendo, Robertson (1981) indica as principais características de um "bom" sistema de recuperação da informação:

- **Efetividade** (significa quão bem ele desempenha uma tarefa delegada);
- **Benefício** (o quanto se ganha com a sua utilização em determinado contexto); e

[10] O autor chamou de ***one-off occurence***.

- **Eficiência** (relaciona-se com o custo de toda a operação, isto é, equilíbrio entre custo e benefício).

Sobre esta problemática, existe ainda a questão dos sistemas informatizados de recuperação da informação, que segundo Rowley (2002), passaram por inúmeras transformações que podem ser condensadas em três estágios de desenvolvimento ou gerações nos últimos vinte anos:

Tabela 3 – Gerações de sistemas de recuperação da informação

Primeira geração	Metadados	Interfaces baseadas em comandos, usuários especialistas e intermediários; número limitado de sistemas em linha nas instituições e disponíveis externamente por meio de serviços de busca em linha
Segunda geração	Dados com texto integral	Interfaces baseadas em menus e comandos; recursos de recuperação adicionais, como hipertexto e buscas em texto completo; interfaces baseadas em DOS; previsto o acesso pelo usuário final, mas nem sempre possível ou alcançado; sistemas em linha, com os primeiros sistemas baseados em cederrom
Terceira geração	Multimídia	Interfaces gráficas; foco no acesso pelo usuário final; orientada para o mercado e com ênfase em pacotes de produtos; armazenamento e distribuição em cederrom ou em redes de alta capacidade; multimídia; intermediário com a função de instrutor; maior uso no lar e em ambientes de acesso público

Fonte: Rowley (2002).

Cada uma das gerações apresentadas na Tabela 3 estava assentada na tecnologia mais avançada de sua época. Deste modo, segundo a autora, os tipos de dados armazenados nos sistemas, a conectividade de sistemas, a interface do usuários e a natureza do grupo de usuários, passaram por mudanças gradativas também.

A fim de ilustrar e complementar a problemática em torno dos sistemas de recuperação da informação, apresentamos as suas funções colocadas em fluxograma proposto por Cianconi (1990) na Figura 16.

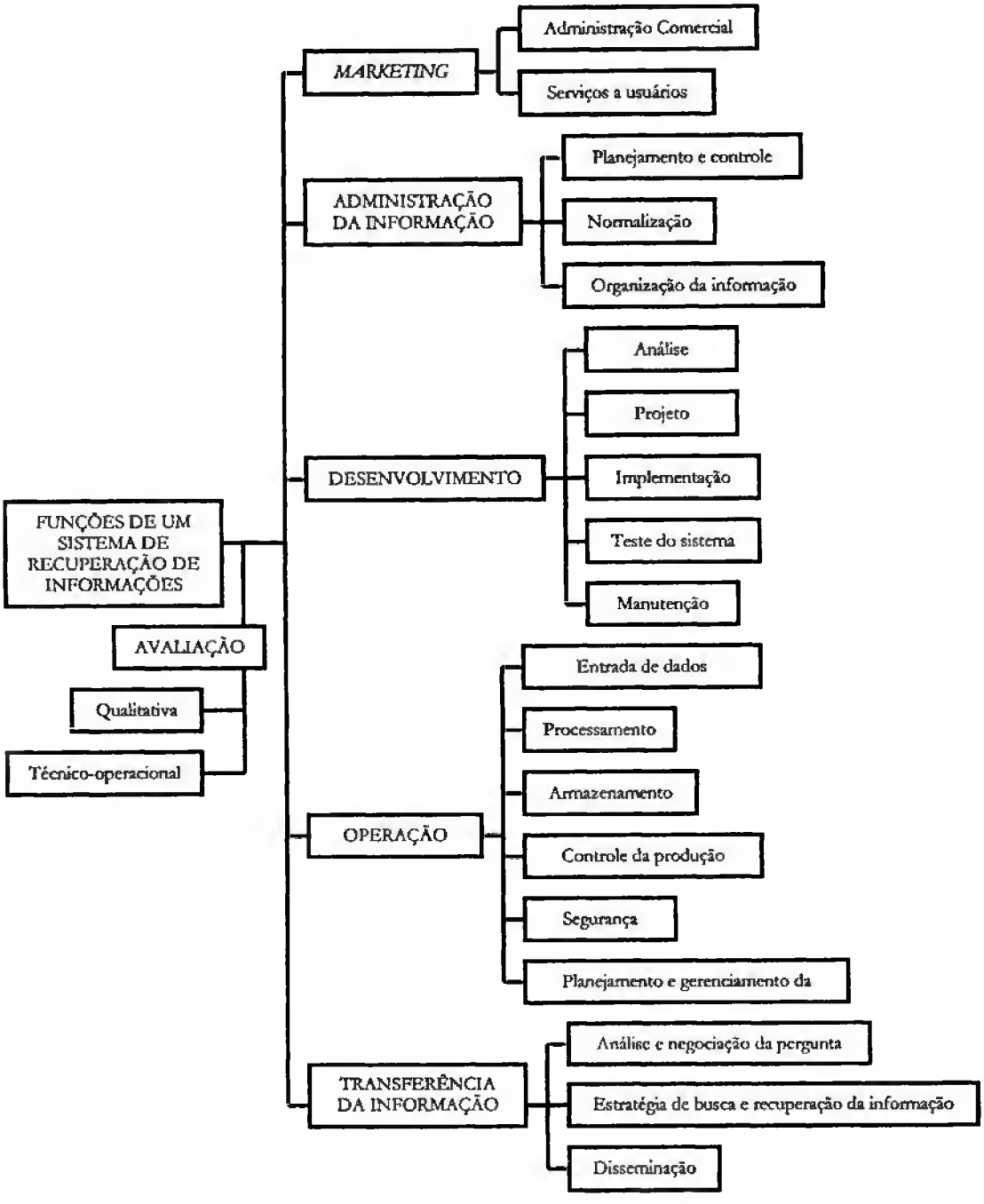


Figura 16 – Funções de um sistema de recuperação de informações
(Fonte: adaptado de Cianconi, 1990)

Retomando a questão dos testes de efetividade de sistemas de recuperação da informação, Belkin (1981) relata que há um grande número de conceitos (fenômenos) que dificultam os testes dos sistemas de recuperação da informação. As dificuldades impostas decorrem primeiramente de sua natureza dúbia (são difíceis de definir ou mesmo de conceituar) e o seu significado é parte central de um processo de recuperação da informação.

Este problema é resumido pelo autor como sendo: "a efetiva transformação da necessidade de informação, da idealização ao uso prático".Com efeito, Belkin (1981) especifica uma situação por meio das seguintes características:

A) O usuário reconhece uma necessidade de informação e a apresenta a um mecanismo de recuperação da informação (coleção de textos, por exemplo), pesquisa e aguarda que o mecanismo de recuperação da informação seja capaz de satisfazer a necessidade;

B) O papel do mecanismo de recuperação da informação é fornecer textos que se aproximem da demanda registrada;

C) O usuário examina partes ou todos os textos apresentados (recuperados) e julga se a sua necessidade está satisfeita, parcialmente satisfeita ou não satisfeita, com base em critérios de utilização e utilidade dos mesmos.

Para Belkin (1981), os conceitos fundamentais com os quais o processo de recuperação da informação lida são: necessidade de informação, desejo, informação, significado ou mesmo falta de significado, satisfação (incluindo precisão), e efetividade (da informação). Tais conceitos são importantes na medida em que são as bases para o funcionamento do processo de recuperação da informação e neste contexto podem ser categorizados da seguinte forma:

1) Conceitos relacionados aos usuários: informação, necessidade e desejo;

2) Conceitos relacionados aos textos: informação, significado, falta de significado; e

3) Conceitos relacionados aos usuários e aos textos: satisfação e efetividade.

Esta categorização é parte integrante da estrutura de um sistema de recuperação da informação, onde há a separação de documentos e necessidade e a tentativa de co-relacionar um com o outro. Geralmente para a realização de testes em sistemas de recuperação da informação, não são consideradas as categorias acima descritas. Para o autor os conceitos apresentados deveriam estar implícitos em todo o sistema cujo objetivo precípuo seja a recuperação da informação.

A relação entre a necessidade de informação do usuário/demandante e as principais características de um bom sistema de recuperação da informação, estão presentes nas argumentações apresentadas por Robertson (1981). Na proposta do autor a efetividade, o benefício e a eficiência impactam diretamente os conceitos relacionados aos usuários: informação, necessidade e desejo. Na Figura abaixo, está proposto um modelo de solução para as necessidades de informação dos usuários adaptado de Checkland (1999):

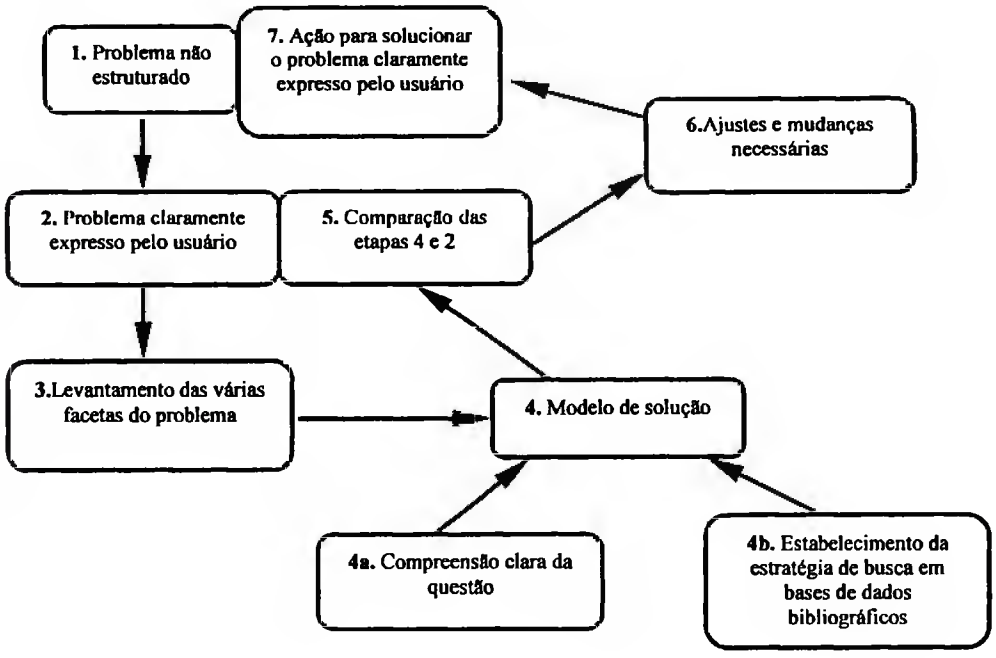


Figura 17 – Modelo de solução para as necessidades de informação dos usuários (Fonte: adaptado de Checkland, 1999)

A representação da Figura 17 guarda semelhanças com as etapas decisórias que consubstanciam o 'processo normal de referência' proposto por Grogan (1995) explicitado na seção anterior do trabalho. As semelhanças anotadas podem ser cotejadas no Quadro a seguir:

Quadro 2 – Semelhanças entre o processo normal de referência e o modelo de solução para as necessidades de informação dos usuários.

Processo normal de referência (proposto por Grogan, 1995)	*Modelo de solução para as necessidades de informação dos usuários (adaptado de Checkland, 1999)*
Etapa 1 – Problema; **Etapa 2** – Necessidade de informação;	**1** – Problema não estruturado;
Etapa 3 – Questão inicial;	**2** – Problema claramente expresso pelo usuário;
Etapa 4 – Questão negociada;	**3** – Levantamento das várias facetas do problema;
Etapa 5 – Estratégia de busca;	**4** – Modelo de solução (4a – Compreensão clara da questão e 4b – Estabelecimento da estratégia de busca em base de dados bibliográficos);
Etapa 6 – Processo de busca;	**5** – Comparação do problema claramente expresso pelo usuário com o modelo de solução;
Etapa 7 – Resposta; e	**6** – Ajustes e mudanças necessárias;
Etapa 8 – Solução.	**7** – Ação para solucionar o problema claramente expresso pelo usuário.

Outros tantos modelos foram propostos para demonstrar a importância que a identificação e a compreensão das necessidades informacionais do usuário representam para o processo de busca e recuperação da informação.

Um destes modelos foi proposto por Wilson (1999):

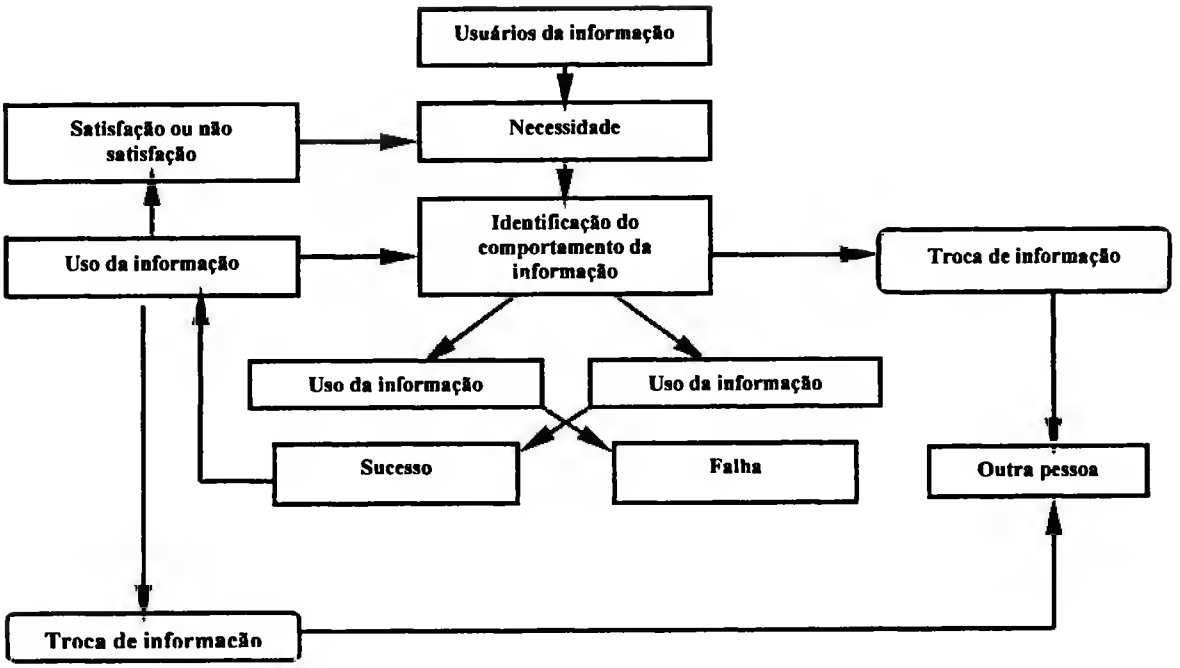


Figura 18 – Modelo de Wilson do comportamento da informação
(Fonte: Wilson, 1999)

As questões conceituais aqui levantadas visam dar uma noção mais clara do caráter fundamental do processo de busca e recuperação da informação nos sistemas de recuperação da informação.

A associação das necessidades de informação, conhecida como insumo à montagem do modelo de busca e recuperação da informação, é decisiva e pode ser demonstrada pela análise dos índices de precisão obtidos. Com isso, a posição da atividade de recuperação da informação presente no processo de referência (processo normal de referência de Grogan) deve estar sincronizada com qualquer modelo que possa ser proposto como solução para as necessidades de informação dos usuários.

Daí a junção proposta entre o processo normal de referência e modelo adaptado de Checkland (1999), para as necessidades de informação dos usuários.

Inúmeros modelos têm sido construídos a fim de demonstrar todos os elementos envolvidos no processo de busca e recuperação da informação, bem como as suas interfaces. Como exemplo, tomamos o modelo desenvolvido por Ingwersen (1996) *apud* Wilson (1999) na Figura 19, onde são introduzidas as idéias do espaço cognitivo e do meio ambiente organizacional/social.

Figura 19 – Modelo do processo de recuperação da informação de Ingwersen (Fonte: Wilson, 1999)

Mais tarde o mesmo Ingwersen (1999) propõe um avanço no Modelo do processo de recuperação da informação no sistema cognitivo de comunicação para a recuperação da informação, representado na Figura 20 a seguir:

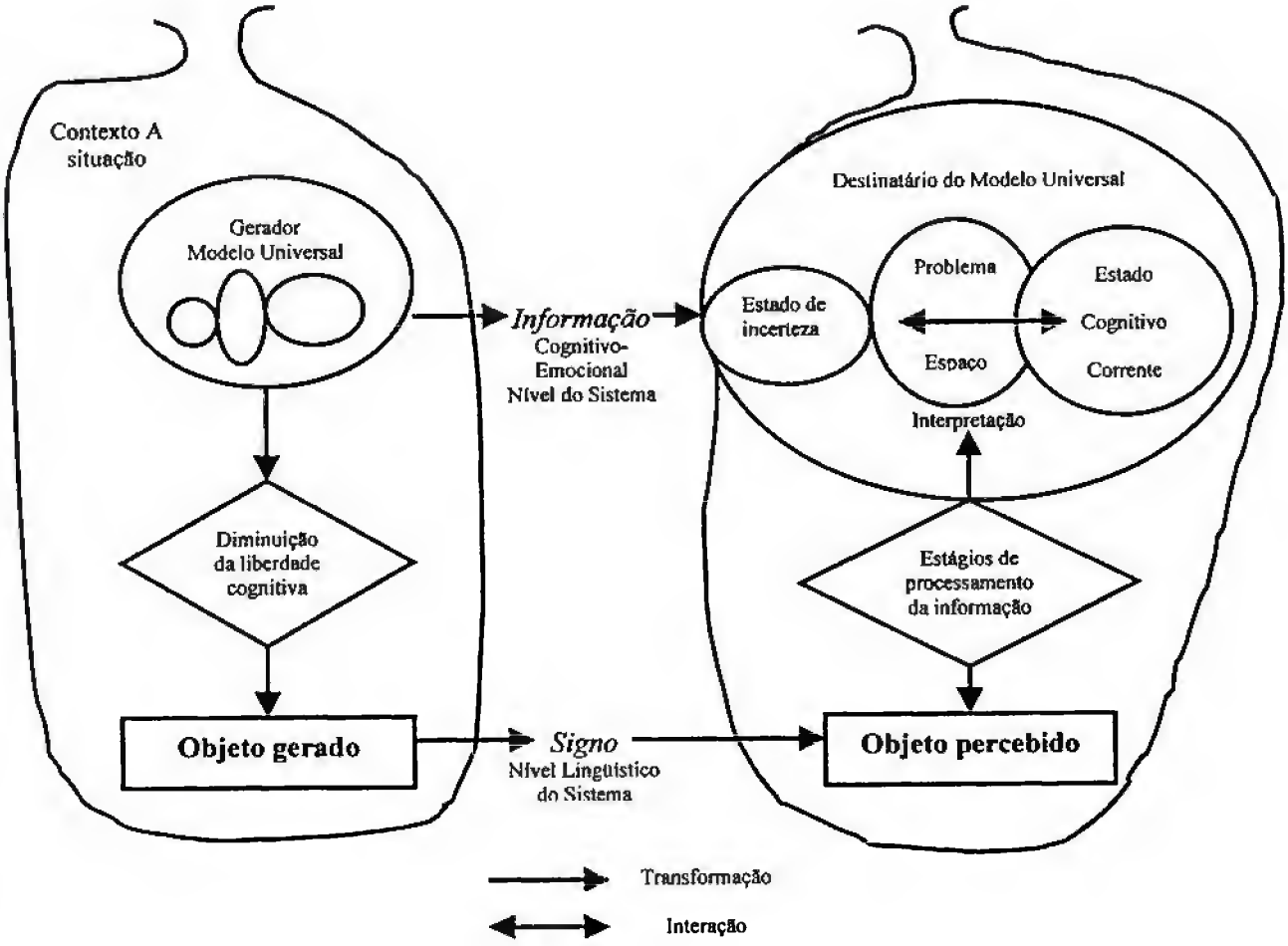


Figura 20 – Sistema cognitivo de comunicação para a recuperação da informação (Fonte: Ingwersen, 1999)

Para Rowley (2002), a recuperação da informação pode ser pensada envolvendo três etapas básicas:

- Aceitação de uma consulta como insumo (como uma representação da necessidade de informação) formulada pelo usuário;
- Execução de uma comparação da consulta com cada um dos registros (representações dos documentos) existentes na base de dados; e
- Produção como resultado, a ser submetido ao usuário, de um conjunto de registros recuperados e que foram identificados com base nesta comparação.

A última das três etapas descritas por Rowley representa o princípio básico do cálculo do índice de precisão, já que envolve a submissão dos resultados da recuperação da informação à apreciação do usuário. A partir daí, o cálculo da precisão se concretiza com o *feedback* do usuário sobre a utilidade de cada item bibliográfico a ele submetido.

O papel da precisão em um processo de busca e recuperação da informação é dar a noção exata se o que está sendo recuperado na base de dados é útil ao usuário. Desta conclusão uma série de decisões poderão mudar os rumos de qualidade da resposta que se obtém nos sistemas de recuperação da informação.

4.3.2- Conclusão

A utilização dos índices de precisão deverá sempre propiciar parâmetros para a melhoria contínua da resposta obtida dos sistemas de recuperação da informação. O conhecimento das necessidades de informação dos usuários é ponto de partida para a concretização desta meta, entretanto, tais necessidades acabam por gerar também graus de imprecisão, tal como observa Foskett (1996), ou seja, incapacidade de um sistema de informação de recuperar documentos úteis frente à solicitação do usuário, sobretudo se envolver a negociação da questão.

Deste modo, a montagem de um modelo de busca e recuperação da informação tem a finalidade de estabelecer uma estratégia que possa minimizar a imprecisão, daí a condição básica de se negociar a questão, que de acordo com Grogan (1995) representa o ajuste necessário à compreensão clara da demanda informacional.

Além do conhecimento das necessidades informacionais do usuário e da montagem do modelo de busca e recuperação da informação, a especificidade e a exaustividade devem ser amplamente consideradas. A ocorrência da

especificidade e da exaustividade se dá pela análise da demanda informacional do usuário, a fim de estabelecer o quanto o aumento da exaustividade é viável sem ampliar a revocação que mede se um item foi ou não recuperado e a extensão desta recuperação. Enfim, a conjugação dos três fatores elencados é condição básica e necessária na busca de índices de precisão mais altos.

Para o uso do índice de precisão (ver item 4.4) como parâmetro para a avaliação da qualidade do processo de busca e recuperação da informação, o fluxograma da Figura 21 reúne uma boa parte dos elementos a serem considerados:

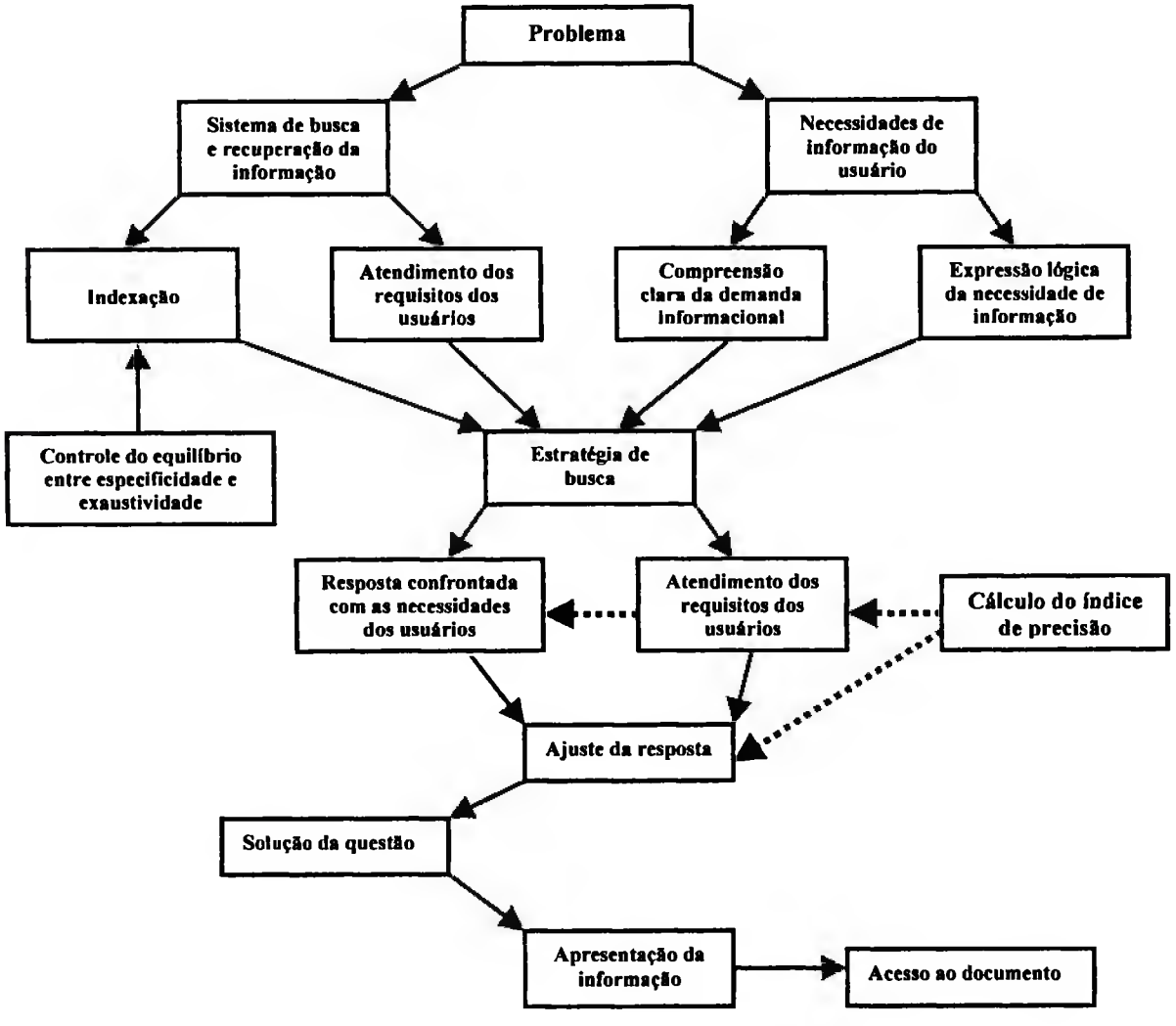


Figura 21 – Fluxograma da incidência do cálculo da precisão no processo de busca e recuperação da informação

4.4- Precisão

A Precisão é um conceito fundamental para a avaliação da qualidade da recuperação da informação, ao mesmo tempo que representa a medida de interesse (informação útil) do que foi encontrado em um processo de busca e recuperação da informação para o usuário, que qualifica a informação recuperada como útil ou inútil de acordo com as suas necessidades.

4.4.1- Conceitos e índice de precisão

A precisão pode ser conceituada, segundo Lancaster (1998), como a extensão com a qual os itens recuperados em um processo de busca e recuperação da informação em uma base de dados são considerados úteis. A alta precisão é dada quando a maioria ou a totalidade dos itens bibliográficos recuperados for considerada útil.

A questão da precisão foi amplamente discutida em associação com a análise de desempenho de um sistema de recuperação da informação. Lancaster & Fayen (1973) já preconizavam que ao considerar os fatores que afetam o desempenho destes sistemas será necessário, antes de tudo, conhecer quais são os pré-requisitos do usuário em relação aos resultados da busca e recuperação da informação.

Ao avançar nas considerações sobre o papel do usuário, os autores listam pré-requisitos de acordo com a sua maior ou menor incidência:

I. Abrangência/amplitude;
II. Revocação;
III. Precisão;
IV. Tempo de resposta;
V. Esforço de busca; e

VI. Formato de saída dos dados.

A precisão pode ser determinada por uma proporção, conclusão a que se chega em pesquisas feitas na literatura corrente: se 50 documentos são recuperados e o usuário julgar que 10 são úteis, o índice de precisão será calculado pela proporção de 10/50, que dará um resultado de 20%.

Quando uma pesquisa é realizada em uma base de dados, o sistema divide os resultados da busca em duas partes:

I. Os documentos que atendem à estratégia de busca (a + b); e
II. Os documentos que não se enquadram na estratégia usada (c + d).

Na maioria das vezes, os documentos recuperados representam um percentual muito pequeno em relação ao tamanho da coleção de dados. Em outras palavras, o total de (a + b) é menor que o total de (c + d), o número de documentos não recuperados é bem grande. Por exemplo: uma pesquisa que recuperou 80 documentos de uma coleção de 500.000 referências, mostra que a + b = 80 e c + d = 499,20 (Lancaster & Warner, 1993).

Os autores propõem a Tabela 4 para ilustrar a equação acima descrita:

Tabela 4 - Indicação de utilidade por parte do usuário

	Úteis	*Inúteis*	*Total*
Recuperadas	a (encontradas)	b (ruído)	a + b
Não recuperadas	c (perdidas)	d (rejeitadas corretamente)	c + d
Total	a + c	b + d	a + b + c + d (coleção total)

Fonte: Lancaster & Warner (1993)

A variação no índice de precisão depende fundamentalmente do julgamento do usuário para com o conjunto de itens recuperados, ou seja, se todos os itens

recuperados na base de dados forem úteis ao usuário, a precisão será total ou de 100% (Heaps, 1978).

Tradicionalmente a precisão é relacionada à revocação em uma relação inversamente proporcional, o que significa dizer que quanto maior for a precisão menor será a revocação.

A revocação pelo clássico estudo de Borko & Bernier (1978) pode ser conceituada como a porcentagem de itens úteis que um termo ou combinação de termos, pode recuperar.

Para Lancaster & Warner (1993), a revocação mede se um item foi ou não recuperado e a extensão da recuperação de itens bibliográficos. No caso de um usuário estar realizando uma busca em determinada coleção, a revocação será dada se houver acesso ao documento necessário, logo após o processo de busca e recuperação da informação. Se há uma busca por termos em uma base de dados, a revocação é medida de acordo com o número de referências úteis recuperadas.

A precisão e a revocação foram usadas pela primeira vez no estudo de Cleverdon, intitulado *Report on testing and analysis of investigation into comparative efficiency of indexing systems* de 1962. Neste trabalho, o autor apresenta os dois parâmetros, revocação e precisão que são definidos por:

$$R = \frac{\text{Número de referências úteis e recuperadas}}{\text{Número de referências relevantes no arquivo}} = \frac{a}{a + b}$$

$$P = \frac{\text{Número de referências úteis e recuperadas}}{\text{Total de referências recuperadas}} = \frac{a}{a + c}$$

Onde:

R = revocação;

P = precisão;

a = referências úteis e recuperadas;

b = referências úteis não recuperadas; e

c = referências inúteis e recuperadas.

Em um segundo estudo, agora com Fayen, Lancaster (1973) propõe que a relação da revocação e da precisão incidem diretamente no nível de assertividade da recuperação da informação em um sistema de busca. Assim, é possível expressar quantitativamente o nível de sucesso da recuperação da informação em um sistema de busca. O índice proposto é o da revocação, que é dado também a partir da fórmula:

$$\frac{\text{Número de documentos úteis encontrados pelo sistema}}{\text{Número total de documentos úteis contidos no sistema}} \text{ X 100}$$

Como exemplo, os autores supõem que um usuário realiza uma busca em um dado sistema e identifica somente 10 itens relevantes. Sete destes itens foram acessados na íntegra pelo usuário. A relação então é: 7/10 X 100 ou 70%.

Mesmo considerando que o Índice de revocação é uma medida válida e importante no sucesso da busca, outra medida deve ser considerada o índice de precisão, já que o sistema de busca pode trazer o que de fato o usuário necessita mas também, o que não é necessário. O índice de precisão também é dado por:

$$\frac{\text{Número de documentos úteis recuperados pelo sistema}}{\text{Número total de documentos encontrados pelo sistema}} \text{ X 100}$$

Assim como Lancaster & Fayen (1973), Foskett (1996), propõe a relação da revocação e a relação da precisão:

$$\text{Relação da revocação} = \frac{A \cap B}{A} = \frac{\text{(documentos úteis recuperados)}}{\text{(total de documentos úteis)}}$$

$$\text{Relação da precisão} = \frac{A \cap B}{B} = \frac{\text{(documentos úteis recuperados)}}{\text{(total de documentos recuperados)}}$$

As porcentagens são obtidas multiplicando as relações por 100.

Do mesmo modo e confirmando a proposição dos outros autores, Dunlop (2000) propõe a Relação da revocação e a Relação da precisão a fim de subsidiar a mensuração de como efetivamente um sistema acha documentos úteis:

$$\text{Revocação} = \frac{\text{Número de documentos úteis recuperados até aqui}}{\text{Número total de documentos úteis}}$$

$$\text{Precisão} = \frac{\text{Número de documentos úteis recuperados até aqui}}{\text{Número total de documentos úteis até aqui}}$$

Mais uma vez, a conclusão que se deduz das proposições formuladas é a de que a revocação e a precisão variam inversamente, ou seja, quando existe incidência de revocação há conseqüentemente redução da precisão e quando há maior precisão haverá redução da revocação. A figura 22 apresenta esquematicamente esta relação:

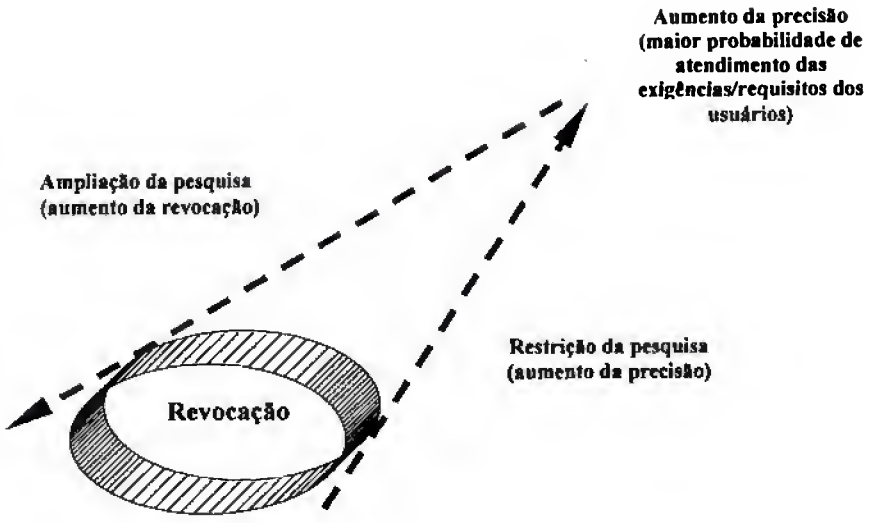


Figura 22 – Revocação e precisão
(Fonte: baseado em Foskett, 1996)

Outros dois conceitos são importantes na compreensão da precisão: a especificidade e a exaustividade. A especificidade diz respeito à precisão que pode ser conseguida ao detalharmos o assunto que estamos buscando. Diante disto, quanto maior a especificidade, tanto maior será a probabilidade de alcance da alta precisão.

Para Foskett (1996), caso seja necessário aumentar a revocação, certamente uma parte da especificidade estará prejudicada, fato que implicará na impossibilidade de aumento da precisão.

A exaustividade, ao contrário da especificidade, é uma decisão que pode ser administrativa, visto que a extensão com que se analisa um dado documento se destina ao estabelecimento do conteúdo temático a ser especificado.

Considerando o objetivo da abordagem da especificidade e da exaustividade na análise da precisão, em princípio fica claro que o aumento da exaustividade pode trazer pouca ou nenhuma vantagem no incremento da

precisão, em oposição à especificidade que poderá trazer benefícios. A Figura 23 ilustra esta relação:

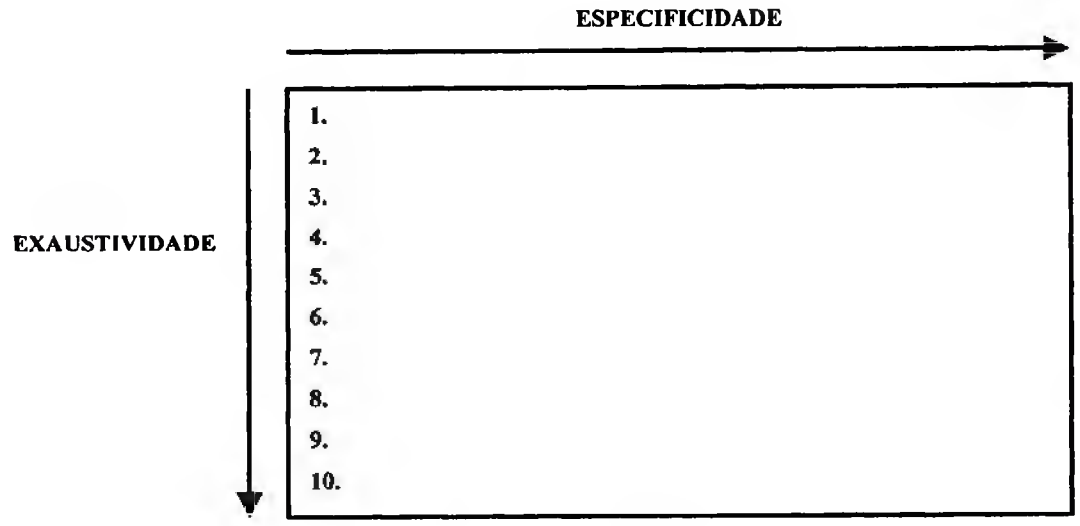


Figura 23 – As duas dimensões da indexação de um documento
(Fonte: Lancaster, 1993)

Não obstante as conclusões apresentadas, o usuário possui um papel-chave na determinação da precisão, inclusive no que diz respeito aos fatores que influem na performance dos sistemas de busca e recuperação da informação. Com isso, a exaustividade e a especificidade deverão ser definidas com base nas necessidades do usuário.

Sobre os principais fatores que contribuem para a efetividade da busca e recuperação da informação em um sistema estão listados a arquitetura do sistema, fatores ligados ao plano diretor de sistemas, sobretudo no que corresponde às soluções de *hardware* e *software*, treinamento para a utilização do sistema e, fundamentalmente, naquilo que Lancaster & Fayen (1973) chamaram de fatores intelectuais relacionados à base de dados (características da indexação, vocabulário e a estratégia de busca necessária para a recuperação da

informação). As principais causas de falhas no processo de pesquisa em sistemas de busca de informação são apresentadas na Tabela 5:

Tabela 5 – Principais causas de falhas no processo de pesquisa em sistemas de busca de informação

	Falha de revocação	*Falha de precisão*
Linguagem de indexação	- Ausência de termos específicos (vocabulário de entrada) - Hierarquia inadequada ou referência cruzada com estrutura inadequada - Indicadores de função, causando grande imprecisão	- Ausência de termos específicos (descritores) - Falhas na hierarquia - Falsas coordenações - Termos relacionados incorretos
Indexação	- Ausência de especificidade - Ausência de exaustividade - Omissão de conceitos importantes - Uso de termos não apropriados	- Indexação exaustiva - Uso de termos não apropriados
Busca	- Falhas na abrangência de todas as possibilidades de recuperação - Estratégias muito exaustivas - Estratégias muito específicas	- Estratégia não suficientemente exaustiva - Estratégia não suficientemente específica - Uso de termos não apropriados ou combinação de termos - Erro na lógica de busca
Interação usuário/sistema	- Recuperação mais específica que a atual necessidade de informação	- Recuperação mais específica que a atual necessidade de informação

Fonte: Lancaster & Fayen (1973)

Outro conceito que também se insere com freqüência no contexto da precisão é a relevância. Na literatura, em vários momentos, os conceitos de precisão e relevância se confundem, entretanto são distintos.

Como forma de dirimir dúvidas sobre esta sobreposição de conceitos, cabe esclarecer o que vem a ser relevância e a discussão que está na sua consideração. Para tanto, foram selecionados três autores que abordaram o tema em questão: Saracevic (1999), Froehlich (1994) e Robredo (2003).

De acordo com Robredo (2003) em seu trabalho intitulado: Da ciência da informação revisitada aos sistemas humanos de informação, o conceito de relevância é caro tanto para a Ciência da Informação como para a Ciência da Computação e pode ser definido como a capacidade de um motor ou de uma função de busca de recuperar dados e informações de fato úteis no atendimento das necessidades de informação dos usuários dos sistemas.

Já a pesquisa de Froehlich (1994) intitulada: *Relevance reconsidered-towards an agenda for 21st century*, pretendeu provar que os sistemas de informação devem ser modelados e desenvolvidos de modo que os usuários possam filtrar e classificar os resultados de suas buscas baseando-se em critérios que eles próprios achem adequados. Todavia, um conhecimento mais aprofundado deve ser buscado para a determinação da relevância, relacionando o usuário com a sua necessidade de informação. Assim sendo, os pesquisadores desta área devem acrescentar aos seus trabalhos, a investigação de estruturas da sociologia do conhecimento e da epistemologia social.

No trabalho: *Information science* de autoria de Saracevic (1999), o autor afirma que como todo fenômeno complexo, o conceito de relevância possui uma trajetória longa e turbulenta no âmbito da recuperação da informação, assumindo definições específicas que variam de acordo com os contextos de aplicação, sendo no caso da Ciência da Informação o atributo ou critério que reflete a efetividade da troca de informações entre usuários e sistemas de recuperação da informação.

O autor considera ainda que a relevância indica uma relação, desta forma, pode-se distinguir nos diferentes tipos de relações, diferentes tipos de relevância:

- **Relevância sistêmica ou algorítmica:** relação entre a pergunta e a informação (textos) contida em arquivos do sistema de recuperação da

informação. A comparação da efetividade na relevância é o critério adotado neste tipo de sistema;

- **Relevância temática ou do assunto:** relação entre o assunto ou tema expresso na pergunta e o tema ou assunto coberto pelos textos recuperados, ou de maneira geral, os textos contidos no sistema. Especificidade é o critério deste tipo de relevância;
- **Relevância cognitiva ou pertinência:** relação entre o estado do conhecimento, a necessidade de informação cognitiva do usuário e os textos recuperados no arquivo do sistema. A correspondência cognitiva, novidade, qualidade e necessidade de informação são os critérios deste tipo de relevância;
- **Relevância situacional ou utilitária:** relação entre a situação, tarefa ou problema prático e os textos recuperados pelo sistema. A utilidade na tomada de decisão, a precisão da informação na resolução de problemas, a redução de incerteza são os critérios deste tipo de relevância; e
- **Relevância motivacional ou afetiva:** relação entre as intenções, objetivos e motivações do usuário e os textos recuperados pelo sistema. A satisfação do usuário é o critério deste tipo de relevância.

Saracevic (1999) enfatiza, mais uma vez, que o objetivo principal dos sistemas de recuperação da informação deve estar calcado na busca da relevância para o usuário. Entretanto, o usuário poderá julgar a relevância sob óticas pessoais dependendo da sua necessidade específica de informação, tornando o conceito da relevância um tanto flexível e como conseqüência, difícil de estudar sob parâmetros mais precisos.

4.4.2- Gestão da precisão

A precisão é uma medida objetiva de rendimento, por isso, a sua gestão deve estar apoiada em dois elementos básicos:

I. Controle da revocação e da exaustividade, a fim de propiciar o aumento nos índices de precisão no processo de busca e recuperação da informação; e

II. Aperfeiçoamento contínuo da interação dos usuários com os sistemas de recuperação da informação (julgamento do usuário), auxiliando a análise das medidas de precisão apoiada em três fatores:

A) Problema claramente expresso pelo usuário (deve ser empregado a fim de aumentar a especificidade);

B) Levantamento das várias facetas do problema (deve ser empregado a fim de evitar a exaustividade); e

C) Modelo de solução composto por: compreensão clara da questão e estabelecimento da estratégia de busca em base de dados bibliográficos (deve ser construído a fim de evitar a revocação).

A gestão da precisão está ligada também ao valor da informação, que na concepção de Taylor (1986) é a questão central do estudo da transferência da informação. É necessário que se obtenha um consenso sobre o conceito de "valor", baseado na crítica e no teste de métodos que permitam a sua quantificação. De acordo com Slamecka (1970) *apud* Taylor (1986), um resultado eventual de um processo de recuperação da informação pode representar parte de uma das mais difíceis medidas da informação: o seu valor. O desenvolvimento de uma medida pragmática para utilidade/valor da informação foi o principal foco da Ciência da Informação nas últimas décadas.

É inquestionável, na concepção de Taylor, o fato de que o valor da informação está baseado no usuário. Em outras palavras, valor da informação não se dá *per se* e sim pela sua utilização. No entanto, informações soltas para que adquiram valor, necessitarão de um contexto, isto inclui o conhecimento do usuário, de suas necessidades e do uso que faz da informação. Esta é a chave para a "agregação de valor à informação", pois a partir do que o usuário necessita

é que se pode identificar a utilidade de certo tipo de informação, as formas preferenciais de recuperação e a facilidade que um sistema de recuperação da informação pode oferecer. Quando esses elementos estão presentes, podendo transformá-los em rotinas, tem-se um sistema de recuperação da informação que será tão efetivo quanto consistentes forem os insumos sobre o ambiente (contexto) onde o mesmo for utilizado.

Este sistema é parte do processo de agregação de valor, onde o usuário será auxiliado nas escolhas que fizer e nos problemas que resolver.

Um sistema de recuperação da informação vai requerer investimento traduzido em tempo, equipamento e conhecimento. Este é o custo do armazenamento da informação. Na avaliação de Taylor (1986) a palavra armazenamento é tida como ambígua. Significa dar ao usuário uma pilha de papéis que provavelmente contenha alguma informação útil para sua necessidade específica? Significa franquear aos usuários meios eficientes para que eles próprios identifiquem as informações? Significa o armazenamento de informações analisadas, avaliadas e interpretadas para o seu emprego em situações específicas? Na verdade, o armazenamento de informações significa tudo isso e depende da interpretação do contexto e das facilidades que o sistema oferece.

Um sistema de informação constitui-se de uma série de processos formais que permite o incremento das possibilidades de identificação das informações de valor em mensagens. O sistema opera em determinados níveis e o valor acrescentado à informação pode ser externo e/ou interno às mensagens; pode ser tangível, tal como um descritor ou intangível como validação de dados. Desta forma, os processos que agregam valor à informação são: **I.** Organização da informação; **II.** Análise; **III.** Decisão e **IV.** Julgamento (Taylor, 1986).

A partir do momento em que a informação recuperada possua características de valor agregado reconhecidas pelos usuários, o seu valor

aumenta em uma relação direta a cada item recuperado pelo sistema passando a ser contabilizado no cálculo do índice de precisão.

Na Figura a seguir estão apresentados os elementos que compõem o espectro do valor agregado.

Figura 24 – Espectro do valor agregado
(Fonte: Taylor, 1986)

Os conceitos de "valor" e "valor agregado" são comumente associados entre si. A primeira interpretação de "valor agregado" é emprestada da área de economia. Basicamente agregar valor é criar "riqueza". O conceito não é novo, foi desenvolvido por Tench Coxe, um economista norte-americano que foi Secretário Assistente do Tesouro em Washington. Segundo ele: "valor agregado é um tipo de riqueza gerada pelo esforço e pela engenhosidade humana. Uma organização

manufatureira compra matéria-prima, componentes, combustível e contrata vários serviços. Ela converte tudo isso em produtos que deverão ser vendidos por um preço maior que o custo total da matéria-prima e dos outros insumos. Isto é a agregação de valor ao material por meio do processo de produção" (Taylor, 1986).

Pode-se tomar como exemplo da agregação de valor à informação, as etapas na geração de conhecimento e inteligência apresentados na Figura 25:

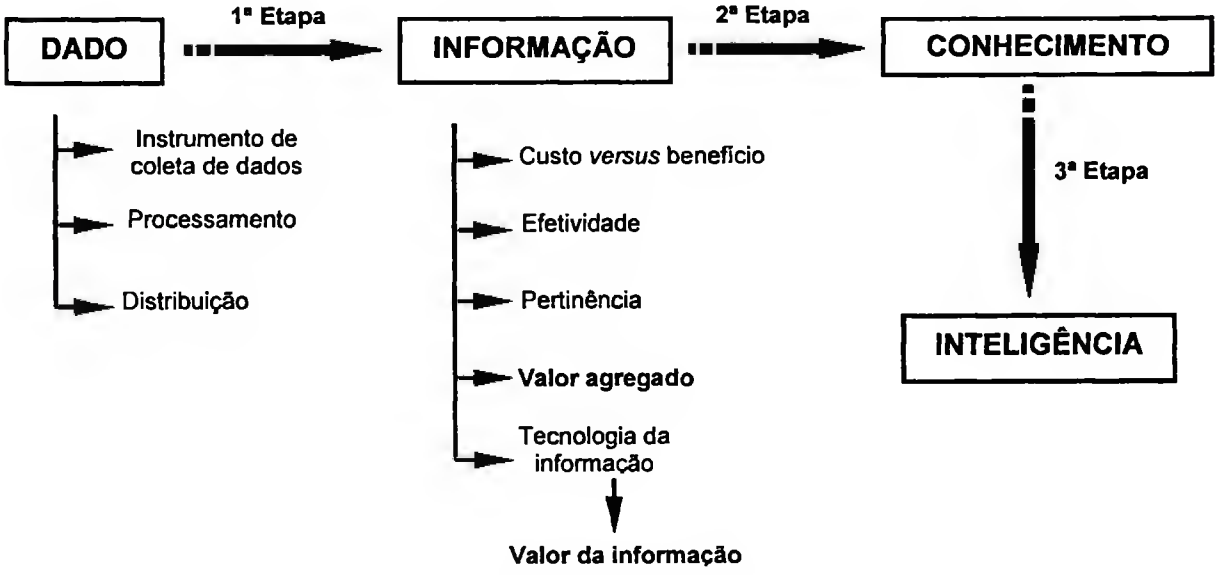


Figura 25 – Etapas na geração de conhecimento e inteligência (Fonte: TARAPANOFF; ARAÚJO Jr. & CORMIER, 2000)

A consideração do usuário como elemento-chave na perspectiva da agregação de valor á informação em seus vários processos, deve ser incluída também, no âmbito da gestão da precisão como ferramenta indispensável na avaliação da resposta que deverá nortear a efetividade do processo de busca e recuperação da informação.

Tal consideração influencia os elementos envolvidos em uma das principais funções do sistema que é a transferência da informação (Figura 26):

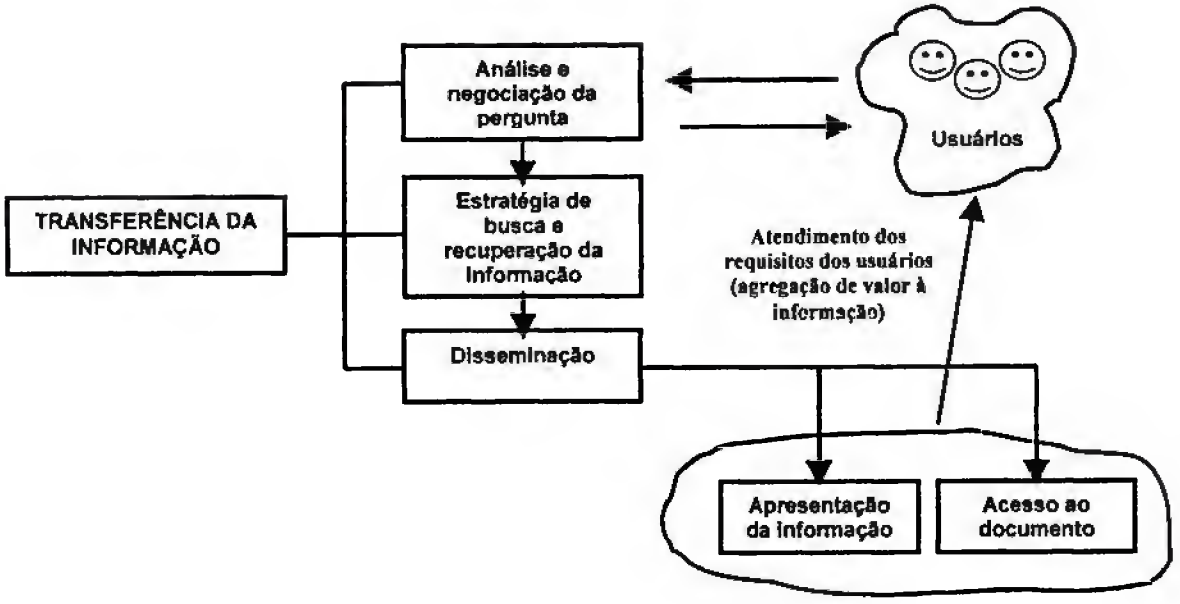


Figura 26 – A perspectiva do usuário na transferência da informação (Fonte: baseado em Cianconi, 1990)

Ao focar a satisfação do usuário como sendo o objetivo final do processo de gerenciamento da informação, o processo de busca e recuperação da informação deverá planejar a ação de incremento da precisão, a partir de tarefas a serem executadas durante todo o processo. A sistematização do gerenciamento estratégico da informação, desta forma, deve ter a seguinte conformidade, como propõem McGee & Prusak (1994):

- Identificação das necessidades e requisitos de informação – tarefa mais importante do processo, pois é neste momento que a decisão de focar as necessidades de informação dos usuários, transforma-se em requisito;
- Classificação e armazenamento de informação/tratamento e apresentação de informação – pressupõe como os usuários terão acesso às informações. Tão importante quanto o conteúdo de cada item informacional, a forma dada por meio do tratamento e da classificação destes itens de informação será decisiva para a posterior recuperação em uma base de dados;

- Desenvolvimento de produtos e serviços de informação – tarefa pela qual os usuários do sistema têm acesso aos itens de informação e ao mesmo tempo possibilitam o cumprimento da tarefa seguinte; e
- Distribuição e disseminação da informação – etapa final do processo, onde os profissionais nela engajados devem estar aptos a compreender com clareza as necessidades de informação dos usuários.

Na concepção proposta pelos autores, a criação de valor para o processo está assentada no conhecimento proativo das necessidades dos usuários, o que deve proporcionar subsídios para a determinação dos requisitos a serem utilizados no âmbito do gerenciamento estratégico da informação.

Os autores colocam que: "embora seja relativamente simples criar um sistema de informações baseado em necessidades predeterminadas, a complexidade do sistema aumenta consideravelmente quando se tenta antecipar essas necessidades. É isso, entretanto, que muitos sistemas de informação devem tentar conseguir, se pretendem alcançar um valor estratégico". A Figura a seguir, fornece uma visão geral de todo o processo:

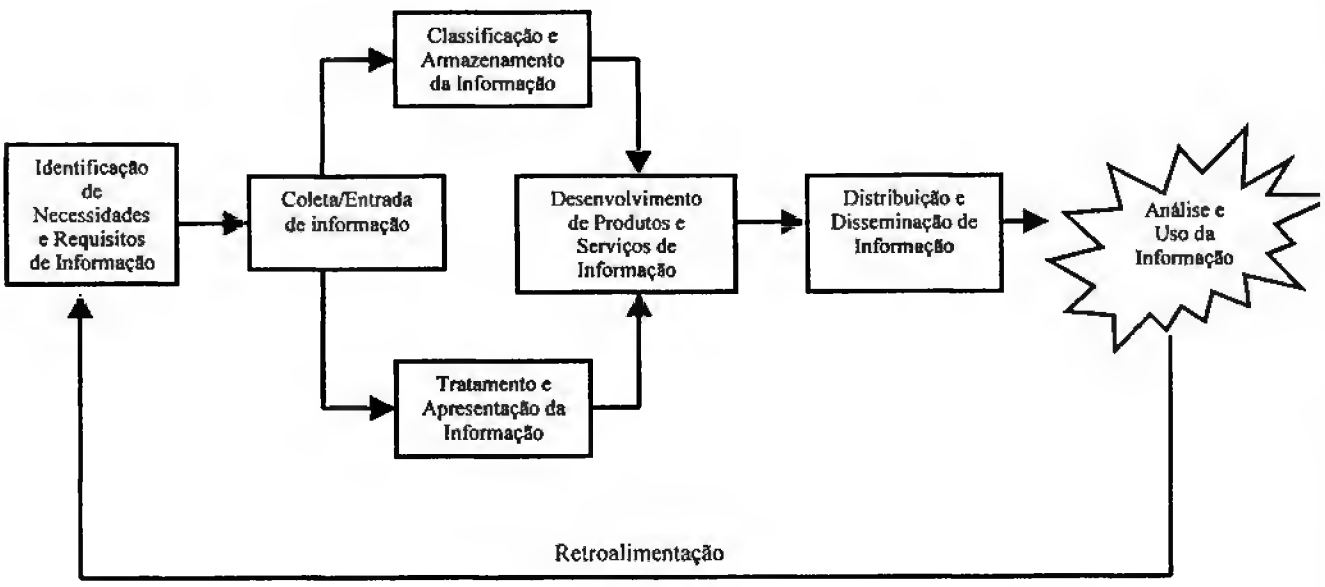


Figura 27 – Tarefas do processo de gerenciamento de informações (Fonte: adaptado de McGee & Prusak, 1994)

O *CRM*, sigla que em inglês significa gerenciamento do relacionamento com o cliente, pode apoiar a busca de foco nas necessidades dos usuários. De acordo com Swift (2000), existem quatro elementos das estratégias do processo de *CRM* que podem ser perfeitamente aplicáveis na relação entre o processo de busca e recuperação da informação e a focalização das necessidades de informação dos usuários.

Os elementos estão representados na Figura abaixo:

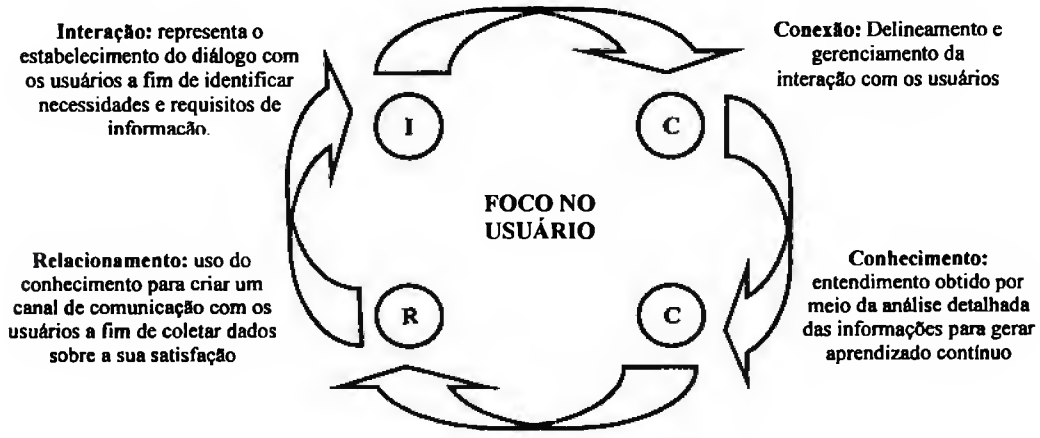


Figura 28 – Estratégias do processo de *CRM* focando o usuário
(Fonte: adaptado de Swift, 2000)

4.4.3- A precisão no processo de busca e recuperação da informação

O conceito de precisão se apresenta como uma medida objetiva de rendimento de sistemas de informação é uma das etapas mais importantes destes sistemas é o resultado do processo de busca e recuperação da informação.

A medida objetiva de rendimento representada pelo índice de precisão permite a aferição do quanto de informação inútil é recuperada, demonstrando a necessidade de detectar em quais componentes do processo de busca e recuperação da informação não há rendimento satisfatório. Decisões que podem ser tomadas a partir de índices de precisão, podem impactar diretamente os

sistemas de informação, ensejando ações para evitar perdas de performance e melhorar a qualidade dos seus serviços.

O problema pode ser detectado em uma análise minuciosa dos índices de precisão obtidos em cada busca efetuada nos sistemas e poderá apontar para falhas em uma das três etapas que formam, segundo Rowley (2002), os sistemas de recuperação da informação:

- Indexação: consiste na atribuição de termos ou códigos a um registro ou documento e que serão posteriormente úteis na recuperação do registro ou documento. A indexação poderá ser intelectual, ou seja, realizada pelo ser humano ou automática, realizada pelo computador;
- Armazenamento: os sistemas de informação se utilizam dos computadores para armazenar os arquivos de documentos e os arquivos de índices, além da manutenção de bases de dados; e
- Recuperação: parte fundamental que depende das etapas de indexação e armazenamento, as quais determinam a melhor estratégia para as buscas em um sistema de recuperação da informação.

A consideração da precisão no âmbito do processo de busca e recuperação da informação implica, efetivamente, no conhecimento das exigências/requisitos dos usuários, para que a sua construção possa estar apoiada não apenas em uma medida objetiva de rendimento da indexação, armazenamento ou recuperação da informação.

Em suma, a problemática da precisão está ligada diretamente ao julgamento do usuário. Julgamento este que deverá ser, por sua vez, colhido a partir de três etapas específicas do processo normal de referência de Grogan, correlacionadas com fatores críticos de sucesso presentes no modelo-solução para as necessidades de informação dos usuários, conforme Figura 17 da página 84, adaptado de Checkland (1999).

Nesta correlação, o problema claramente expresso pelo usuário estará associado diretamente à questão inicial, tendo como fator crítico de sucesso o aumento da especificidade. Já o levantamento das várias facetas do problema liga-se à questão negociada, ensejando precaução no uso da exaustividade. Finalmente, o modelo de solução que inclui a compreensão clara da questão e o estabelecimento da estratégia de busca em base de dados bibliográficos, deve ser formatado a fim de evitar a revocação. O Quadro 3 apresenta a correlação:

Quadro 3 – FCS na correlação entre o processo normal de referência e o modelo de solução para as necessidades de informação dos usuários.

Processo normal de referência	*Modelo de solução para as necessidades de informação dos usuários (Fatores críticos de sucesso)*
Etapa 3 – Questão inicial;	2 – Problema claramente expresso pelo usuário;
Etapa 4 – Questão negociada;	3 – Levantamento das várias facetas do problema (deve ser empregado a fim de evitar a revocação);
Etapa 5 – Estratégia de busca	4 – Modelo de solução composto por: 4a – Compreensão clara da questão e 4b – Estabelecimento da estratégia de busca em base de dados bibliográficos (deve ser construído a fim de evitar a revocação)

Inúmeros fatores influenciam a determinação do que seja útil no julgamento realizado pelo usuário. Como exemplo, imaginemos que um estudante desenvolva uma pesquisa a partir de uma demanda direcionada por um professor. Os critérios de julgamento da utilidade que serão usados pelo estudante estarão baseados na concepção do professor. Isto se dará em detrimento do seu próprio julgamento do que venha a ser utilidade. Em um segundo estágio, se o estudante empreende pesquisa a partir da construção do seu próprio julgamento, este certamente estaria baseado em outras premissas, bem diferentes das orientações do professor (Cole, 2001).

Enfim, uma medida objetiva de rendimento de componentes de um sistema de recuperação da informação, como é o caso da precisão, é imprescindível para

detectar e corrigir as falhas em um processo de recuperação da informação, além de servir de parâmetro para testes de novas tecnologias associadas a melhoria contínua dos processos.

4.4.4- A mineração de textos e o índice de precisão

A mineração de textos é útil na descoberta de padrões inesperados nos textos. Todavia, esta tarefa vai necessitar de uma estrutura organizacional que compreenda o valor que vai ser derivado da informação (Inmon; Terdeman & Imhoff, 2000).

A mineração de textos fará sentido se aplicada a uma situação concreta, onde seja possível sua utilização para verificar o ganho de precisão no processo de busca e recuperação da informação, contribuindo efetivamente para o aprimoramento e busca de valor adicional em todo o processo.

Considerando que o conhecimento dos pré-requisitos dos usuários de um sistema de recuperação da informação é condição necessária para que se saiba quanto de exaustividade e especificidade é necessário para que os índices de precisão sejam altos, a mineração de textos poderá concorrer ou até mesmo apoiar o processo de indexação manual.

Com isso, o aprimoramento da representação temática do conteúdo dos documentos poderá não apenas depender da indexação manual, mas também, do apoio que a ferramenta de mineração de textos poderá trazer, extraindo automaticamente de uma base textual uma lista com as palavras que mais ocorrem neste ou naquele documento, enriquecendo o tesauro. A utilidade desta proposta poderá ser avaliada, *a posteriori*, pela proporção entre o número de documentos úteis recuperados pelo sistema e o número total de documentos encontrados pelo sistema, ou seja, o cálculo do índice de precisão.

Não obstante, para que o aprimoramento da representação temática do conteúdo dos documentos seja contínuo e sistemático, o resultado obtido com o cálculo do índice precisão necessitará ser utilizado como indicador de qualidade dos sistemas de recuperação da informação, inclusive na análise da utilidade da mineração de textos com relação a ganhos de precisão no processo de busca e recuperação da informação.

Com isto, as tarefas de encontrar informações nos textos, tratá-las para que possam ser de fato úteis ao processo de indexação e compará-las com os requisitos informacionais dos usuários, vai possibilitar uma melhor gestão do processo de busca e recuperação da informação, que a utilização do cálculo do índice de precisão poderá confirmar.

4.4.5- Conclusão

A precisão, no âmbito da Ciência da Informação, além de um conceito importante, pode traduzir a própria efetividade dos sistemas de busca e recuperação da informação. Normalmente, as falhas na recuperação da informação podem ser traduzidas como respostas inúteis para os usuários, o que se reflete imediatamente no baixo índice de precisão e repostas úteis, que inclusive interferem na qualidade de um sistema. Neste momento entram os conceitos de revocação e especificidade que descrevem os problemas relativos à restrição ou ampliação da pesquisa, decisões que só fazem sentido se amparadas naquele que é imprescindível para a precisão, o usuário.

Os autores estudados concordam com o papel preponderante que possui o julgamento dos usuários para o cálculo do índice de precisão. Com este consenso se torna factível pensar na precisão como um elemento importante de análise e decisão na busca da melhor resposta nos sistemas de busca e recuperação da informação. Não é possível calcular o número de documentos úteis encontrados pelo sistema sobre o número total de documentos encontrados pelo sistema

multiplicado por 100, sem o julgamento do usuário que demanda tais documentos, ou seja, a precisão se consubstancia por meio de um julgamento externo ao sistema de busca e recuperação da informação, o que vai determinar também, a sua capacidade de atendimento ou o seu desempenho.

Cabe ressaltar ainda, que a precisão não se dá *per se*, mas no contexto em que operam a revocação, a exaustividade a especificidade e sobretudo, tendo como ponto de equilíbrio, o usuário que vai definir, em nome da sua necessidade de informação, o que é útil ou inútil dentre toda a informação recuperada.

4.5- Conclusões da revisão de literatura

As preocupações recorrentes de organizações, usuários e dos profissionais da informação com a melhoria das respostas obtidas em sistemas de recuperação da informação tem sido coincidentes com as preocupações da própria Ciência da Informação.

A finalidade do processo de busca e recuperação da informação em localizar documentos e itens de informação armazenados, só poderá ser validada por intermédio da avaliação dos usuários. Isto significa dizer que os sistemas de recuperação da informação, além de buscar atender às demandas informacionais dos usuários, dependem destes para que a qualidade dos seus serviços seja reconhecida.

Segundo Le Coadic (1994), a recuperação da informação possui limitações associadas à necessidade de informação, fato que acaba por gerar problemas na recuperação de informações úteis às demandas dos usuários. Por conta disto, o investimento que tem sido alocado no desenvolvimento de tecnologias da informação objetivam apresentar novas ferramentas que auxiliem organizações e profissionais na melhoria contínua do desempenho destes sistemas, já que a

informação contida nas bases de dados ou em bases textuais constituem-se em uma das maiores fontes de conhecimento das corporações.

Neste contexto, cabe definir claramente que a necessidade de um indicador para medir o desempenho dos sistemas de recuperação da informação, faz-se necessário. Tal indicador deverá fornecer medidas objetivas sobre o grau de atendimento da demanda informacional dos usuários.

Com o estudo da literatura empreendido, a precisão é um indicador eficaz para traduzir a efetividade dos sistemas de busca e recuperação da informação. Este papel não pode ser destinado à relevância, pois se trata de um termo, que de acordo com Saracevic (1999), possui uma trajetória longa e turbulenta no âmbito da recuperação da informação, assumindo definições específicas que variam de acordo com os contextos de aplicação.

As falhas na recuperação da informação são comumente traduzidas como respostas inúteis para os usuários, situação que interfere diretamente no baixo índice de precisão que é dado pela relação entre número de documentos úteis recuperados e o número total de documentos encontrados pelos sistema.

Outra questão que deve ser considerada é que os diversos autores estudados apontam para o papel preponderante que possui o julgamento dos usuários no cálculo do índice de precisão, pois o número de documentos úteis encontrados pelo sistema será dado por eles. Com base nesta convergência é possível considerar a precisão e especialmente o cálculo do índice de precisão, como elementos importantes de análise e decisão na busca da melhor resposta nos sistemas de busca e recuperação da informação. A Figura 21 da página 90, reúne uma boa parte dos elementos a serem considerados na utilização da precisão como parâmetro para a avaliação da qualidade nos sistemas de recuperação da informação.

Outro elemento importante desta revisão de literatura é o processo de indexação que, por estar associado à descrição e representação do conteúdo dos documentos, exerce uma função considerável no processo de busca e recuperação da informação. A indexação é das tarefas que concorrem na montagem e disponibilização de uma base de dados que cumpra os requisitos mínimos para servir a um processo de recuperação da informação, além do armazenamento e da recuperação propriamente dita.

A tarefa da representação do conteúdo dos documentos determinará os rótulos que cada item de informação receberá ao ser armazenado, pois por este mesmo rótulo as informações deverão ser recuperadas.

Como tarefa associada, o vocabulário do sistema, ou seja, o vocabulário controlado onde os termos empregados na indexação poderão formar um tesauro, desde que exista uma estrutura semântica, vai ser determinante na correta representação do conteúdos dos documentos, além de influenciar as estratégias de busca a serem usadas pelos usuários na recuperação da informação. Pode-se afirmar que um processo de indexação que não incorpore com efetividade suas tarefas poderá ser o responsável por baixos índices de precisão em sistema de recuperação da informação.

Como reflexo desta questão, o estudo de Wellisch (1995), frisa que uma boa parte do sucesso do processo de indexação está nas mãos do indexador, daí a sua proposta de itens indispensáveis que o indexador deve observar, dentre os quais destacam-se:

I. Identificar tópicos e assuntos dentro do texto que venham ao encontro das necessidades dos possíveis usuários daquele índice;

II. Separar tópicos que contenham informações diferentes daquelas necessárias aos usuários;

III. Excluir tópicos que contenham informações diferentes daquelas necessárias aos usuários;

IV. Analisar conceitos, tópicos e assuntos considerados importantes, a fim de que seja providenciada uma lista de entradas para as mesmas;

V. Produzir cabeçalhos que empreguem terminologia usada no documento para auxiliar o processo de recuperação da informação; e

VI. Certificar-se de que o cabeçalho utilizado é apropriado para as necessidades dos usuários e que vai apoiá-los na: **a)** recuperação rápida de alguma informação contida em um documento; **b)** identificação rápida da presença ou ausência de informação em um documento; e **c)** identificação de documentos em uma coleção; entre outros.

Como pode ser constatado nas exigências que incidem sobre o indexador, demonstradas por Wellisch (1995), a indexação é basicamente realizada manualmente (indexação manual ou intelectual), situação que expõe todo o processo a infindáveis erros que comprometem o índice de precisão de resposta nos sistemas de recuperação da informação. Apesar de conceitualmente a indexação automática existir, as tarefas de análise da informação que vão culminar na escolha do termo mais apropriado para representar o conteúdos dos documentos, não pode ser feita totalmente por extração automática ou atribuição automática de termos.

Outro óbice que se interpõe entre a indexação manual e a indexação automática está na análise documentária que passa a ser ferramenta indispensável ao processo de indexação e, para que ela possa cumprir os seus objetivos de auxiliar a representação dos conteúdos dos documentos, será necessário, ainda, considerar nas suas fases de análise e síntese, as necessidades de informação dos usuários do sistema. Infelizmente não existem processos realizados pelo computador, que cumpram estes requisitos.

O outro aspecto da revisão de literatura que é caro ao trabalho de pesquisa apresentado trata da mineração de textos. Esta ferramenta consiste na extração de informações sobre tendências ou padrões em grande volumes de documentos textuais. Uma amostra significativa de informações é avaliada em textos contidos em bases de dados e em fontes de informação em linha (Polanco & François, 2000).

A mineração de textos poderá ser aplicada ao processo de busca e recuperação da informação, como uma ferramenta de extração automática de termos dos documentos contidos nas bases textuais, com a finalidade de compor listas de palavras que mais ocorrem nos documentos a serem analisadas e validadas por indexadores. Com isto a lista de palavras poderá ser usada para enriquecer uma linguagem documentária já disponível sob a forma de linguagem de indexação. Com isso fica mais clara a associação da mineração de textos com o processo de indexação e a sua inserção no processo de busca e recuperação da informação.

Para que esta possibilidade seja viável, a seleção de dados, extração de termos, agrupamento de dados, mapeamento dos agrupamentos, visualização dos resultados e interpretação, deverão ser realizados de modo a gerar instrumentos úteis para o processo de indexação, por isso a opção por duas técnicas de descoberta:

I. Descoberta por listas de conceitos-chave que consiste na apresentação de uma lista contendo os conceitos principais de um dado texto; e

II. Descoberta por descrição de classes de textos que é realizada por meio de uma classe de documentos textuais já agrupados e um tema ou mesmo assunto associado a esta classe (Wives & Loh, 2000).

Na Figura a seguir é apresentada a posição dos dois tipos de descobertas no contexto geral adaptada do processo proposto por Trybula (1999):

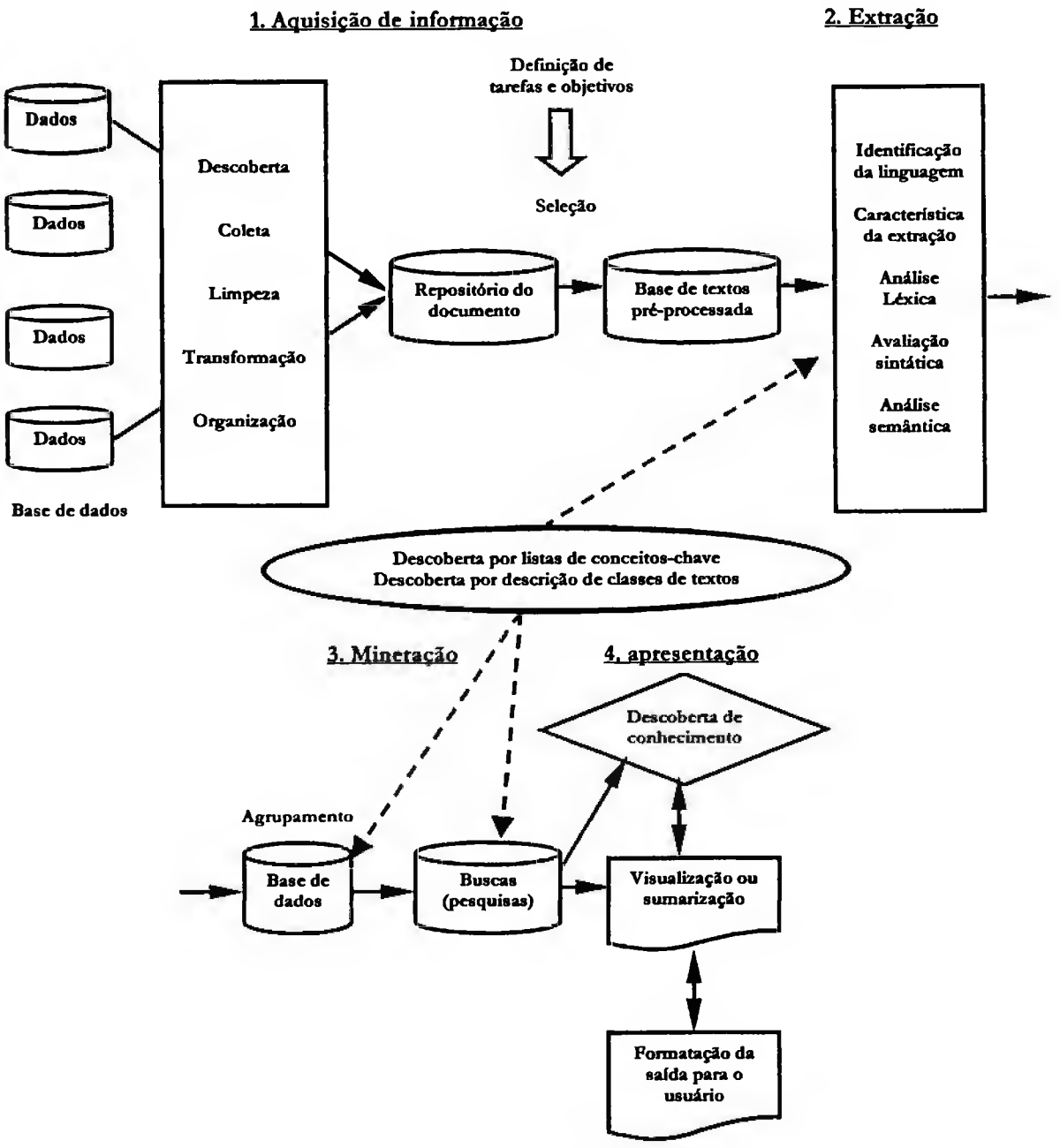


Figura 29 – Incidência das descobertas por listas de conceitos-chave e por descrição de classes de textos no processo de mineração de textos (Fonte: adaptado de Trybula, 1999)

A revisão de literatura procurou abranger e discutir todos os aspectos que envolvem a pesquisa objetivando traçar, de modo claro e coerente, os seus limites. Com isso, teses, pressupostos e variáveis que serão apresentados na seção 5 a seguir, ficam melhor contextualizados no desenvolvimento da pesquisa.

5- TESES, PRESSUPOSTOS E VARIÁVEIS

5.1- Teses

Considerando o problema (conforme explicitado na seção 1.1.2), e baseados na revisão de literatura, defendemos as seguintes teses:

I. O valor do índice de precisão resultante do processo de busca e recuperação da informação baseado na indexação manual, não é superado pelo valor correspondente ao uso da ferramenta de mineração de textos;

II. A mineração de textos pode ser considerada ferramenta de indexação automática por extração automática de termos, desde que seja incluído neste processo o julgamento dos indexadores que deverão selecionar os termos a serem usados na representação do conteúdo dos documentos; e

III. A mineração de textos, desde que entendida como ferramenta de indexação automática, pode ser considerada um instrumento de apoio na construção e/ou enriquecimento do vocabulário controlado, que é gerado e utilizado na indexação manual.

5.2- Pressupostos

1º- Independentemente da utilização de ferramenta de mineração de textos ou da lista de palavras-chave utilizadas na indexação manual, não há ganho significativo no índice de precisão quando do processo de busca e recuperação da informação;

2º- O uso de ferramenta de mineração de textos na busca e recuperação da informação trará sempre como resposta uma maior quantidade de itens bibliográficos do que a lista de palavras-chave utilizadas na indexação manual;

3º- Quando termos específicos da Base de dados do Infohab são submetidos à lista de palavras-chave utilizadas na indexação manual na busca e recuperação da informação, nem sempre são encontrados itens bibliográficos, enquanto que os mesmos termos quando submetidos ao Protótipo com aplicação de mineração de textos, sempre irão recuperar algum item bibliográfico;

5.3- Variáveis

Variáveis do 1º Pressuposto:

A) Índice médio percentual de precisão calculado a partir da média entre os índices resultantes da lista de palavras-chave utilizadas na indexação manual;

B) Índice médio percentual de precisão calculado a partir da média entre os índices resultantes de uso da ferramenta de mineração de textos; e

C) Comparação entre os índices médios percentuais de precisão com base nos resultados obtidos da lista de palavras-chave utilizadas na indexação manual e o uso de ferramenta de mineração de textos.

Variáveis do 2º Pressuposto:

A) Cálculo do número total de itens bibliográficos recuperados da lista de palavras-chave utilizadas na indexação manual;

B) Cálculo do número total de itens bibliográficos recuperados com uso de ferramenta de mineração de textos; e

C) Comparação dos números totais de itens bibliográficos recuperados da lista de palavras-chave utilizadas na indexação manual e do uso de ferramenta de mineração de textos.

Variáveis do 3º Pressuposto:

A) Resultados percentuais nulos obtidos na contagem de itens bibliográficos recuperados na lista de palavras-chave utilizadas na indexação manual; e

B) Resultados percentuais nulos obtidos na contagem de itens bibliográficos recuperados com a utilização de ferramenta de mineração de textos.

Para auxiliar o pleno entendimento dos pressupostos e variáveis a serem exploradas, são apresentadas a seguir algumas definições operacionais úteis à compreensão dos conceitos abordados.

5.4- Definições operacionais

I. **Descoberta por descrição de classes de textos** – Por meio de uma classe de documentos textuais já agrupados e um tema ou mesmo assunto associado a esta classe, a descoberta por descrição deverá facilitar a descoberta das características principais desta classe, a fim de que se possa identificá-la para os usuários e diferenciá-la das outras. Este tipo de descoberta se diferencia da descoberta por listas de conceitos-chave, por descobrir características comuns em vários agrupamentos de textos e não em apenas um texto (Wives & Loh, 2000);

II. **Descoberta por listas de conceitos-chave** - consiste na apresentação de uma lista contendo os conceitos principais de um dado texto. Neste tipo de descoberta o significado do texto é determinado por uma análise de palavras-chave mais importantes, ao invés de uma simples leitura linear (Moscarola, 1998 *apud* Wives & Loh, 2000). Na identificação das palavras-chave podem ser usadas técnicas simples de extração de termos mais freqüentes, por meio da mineração de textos;

III. **Indexação automática** – qualquer procedimento que permita identificar e selecionar os termos que representam o conteúdo dos documentos, sem a

intervenção direta do documentalista. No processo de indexação automática, um algoritmo (ou seja, um conjunto de operações elementares, organizadas logicamente) realiza, em certa medida, o trabalho do indexador no processo de escolha dos termos significativos (Robredo & Cunha, 1986).

IV. Indexação manual – tradução de um documento em termos documentários (descritores, cabeçalhos de assunto e termos-chave) sem o auxílio da atribuição automática de termos ou extração automática de termos. Indexação sem o auxílio de computadores (Rowley, 2002).

V. Indexação por extração automática de termos – consiste na retirada do texto de palavras que serão usadas para representar o seu conteúdo. Indexadores responsáveis por este trabalho deverão selecionar termos e/ou expressões que existem no texto para representar o seu conteúdo temático. Nesta tarefa são naturalmente escolhidas palavras que com maior freqüência ocorrem no texto, sua posição aparecendo no título ou no resumo e o contexto em que aparecem (Lancaster, 1993).

VI. Índice de precisão – índice que mede em termos percentuais a extensão com a qual os itens recuperados em um processo de busca e recuperação da informação em uma base de dados são considerados úteis. O índice de precisão é dado pela seguinte proporção (Cleverdon, 1963). Ver também item 4.4:

$$P = \frac{\textit{Número de documentos úteis recuperados pelo sistema}}{\textit{Número total de documentos encontrados pelo sistema}} \; X\,100$$

VII. Índice médio percentual de precisão – resultado do cálculo da média do resultado de dois ou mais índices de precisão. Considerando uma variável X com observações representadas por $x_1, x_2, \ldots x_n$, a média deste conjunto é a soma dos

valores dividida pelo número total de observações. Assim, o índice médio é dado por (adaptado de Cleverdon, 1963):

$$x = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \times 100$$

VIII. **Item bibliográfico** – representa, por meio de uma referência bibliográfica, cada documento contido em uma base de dados.

IX. **Lista de palavras-chave utilizadas na indexação manual** – Base de dados do Infohab que tem os seus documentos indexados manualmente com o apoio de uma lista de palavras-chave.

X. **Metadados** – enquanto o dado descreve um objeto individual ou entidade e pode ser agrupado para fornecer informação sobre uma população, os metadados provêem informação sobre este dado. Metadados incluem fatos sobre o número e tipos de dados armazenados, tais como a localização e a organização dos mesmos (Delmater & Hancock, 2001).

XI. **Protótipo com uso da ferramenta de mineração de textos** – Base de dados com texto completo de 56 itens bibliográficos criada com o uso da ferramenta de mineração de textos e que correspondem às teses e dissertações extraídas da Base de dados do Infohab.

XII. **Testes de precisão** – procedimento metodológico central da pesquisa, realizado por meio da distribuição e preenchimento de 44 formulários com 22 usuários selecionados dentre os especialistas da CAIXA em ambiente construído. Neste teste as mesmas palavras-chave foram submetidas de forma idêntica ao protótipo e à lista de palavras-chave utilizadas na indexação manual. O resultado

das pesquisas nas bases foi encaminhado para a apreciação e validação destes usuários com a seguinte anotação para cada item bibliográfico recuperado: documento útil (U) ou documento inútil (I). O resultado foi apresentado em duas listas distintas, uma com o resultado da pesquisa na base do Infohab e a outra lista com o resultado da pesquisa no protótipo. Este procedimento visou garantir a correta aplicação da fórmula do cálculo do índice de precisão para cada uma das bases.

6- METODOLOGIA

6.1- Delimitação do estudo

O presente trabalho propôs-se a realizar um estudo comparativo entre a utilidade da ferramenta de mineração de textos e a lista de palavras-chave utilizadas na indexação manual, verificando a variação no índice de precisão durante o processo de busca e recuperação da informação. A investigação teve como universo 1520 documentos da Caixa Econômica Federal (CAIXA) inseridos na base de dados do Centro de Referência e Informação em Habitação (Infohab), empreendimento liderado pela Associação Nacional de Tecnologia do Ambiente Construído (ANTAC) e, como amostra, o acervo de 56 teses e dissertações inseridas no Infohab pela Centralizadora de Documentação e Informação (CEDIN), vinculada à Gerência Nacional de *Marketing* Interno da CAIXA (GEMAC).

O foco do estudo esteve centrado na análise de uma das competências do Infohab, a manutenção de uma base de dados atualizada com referências da produção intelectual dos empregados especialistas em ambiente construído da CAIXA, legislação federal, estadual e municipal, normas pertinentes ao ambiente construído e levantamentos governamentais.

6.2- Caracterização do universo estudado

O universo estudado é representado pelos 1520 documentos da base de dados do Infohab na CAIXA.

O Infohab é um projeto liderado pela Associação Nacional de Tecnologia do Ambiente Construído - ANTAC, sendo formalizado por meio de convênio com Universidades que operam os chamados núcleos do Infohab. De início, sete Universidades Federais participam como núcleos: Universidade Federal Fluminense – UFF; Universidade Federal do Rio de Janeiro – UFRJ; Universidade

de São Paulo – USP; Universidade Federal de Santa Catarina – UFSC; Universidade Federal do Rio Grande do Sul – UFRGS; Universidade Federal da Bahia – UFBA; e Universidade Federal de São Carlos – UFSCAR (Infohab, 2000).

A finalidade do Centro de Referência e Informação em Habitação – Infohab está em captar, selecionar, organizar e divulgar toda a informação da tecnologia do ambiente construído, englobando a sua produção, manutenção e uso, sobretudo no que se refere à habitação de interesse social.

Trata-se de um universo importante no segmento da tecnologia do ambiente construído, com caráter representativo, permanente e de abrangência nacional, ou seja, justificável do ponto de vista da significação da pesquisa ora empreendida, e de onde é possível extrair-se um subconjunto (amostra) de valores parecidos com os do universo que lhe servirá de origem.

No presente caso, o ambiente construído envolve de maneira global todas as atividades, recursos, conhecimento, expertises, experiências, tecnologia, equipamentos, instrumentos, mão-de-obra e mercado relacionados à habitação.

A estrutura organizacional do Infohab é composta por um Comitê Consultivo que possui como atribuições a articulação com outras instituições e a formulação de diretrizes para o planejamento estratégico. Este Comitê é constituído por um representante da Associação Nacional de Tecnologia do Ambiente Construído – ANTAC, o Coordenador geral do Infohab, um representante da Financiadora de Estudos e Projetos – FINEP, órgão do Ministério da Ciência e Tecnologia – MCT, e até seis representantes dos órgãos de fomento e de entidades do setor convidados pelo Fórum dos Coordenadores.

Também faz parte da estrutura organizacional do Infohab um Fórum de Coordenadores que tem como competências:

a) Referendar a escolha de coordenadores e a criação de novos núcleos e grupos associados;

b) Aprovar e extinguir núcleos e grupos associados;

c) Dirigir as atividades, zelando pelo desempenho das tarefas necessárias ao cumprimento dos objetivos do Infohab;

d) Estabelecer o planejamento estratégico, operacional e financeiro do Infohab;

e) Definir a destinação das receitas e aprovar as despesas e balancetes; e

f) Supervisionar e coordenar os núcleos.

O Fórum de Coordenadores tem na sua composição, os coordenadores dos núcleos e o representante da ANTAC.

O Infohab conta também na sua estrutura com a Secretaria Executiva que possui as seguintes competências:

a) Estabelecer e aprovar as parcerias e associações com outras entidades voltadas para a produção, divulgação e tratamento da informação;

b) Gestão financeira em comum acordo com a ANTAC;

c) Ordenamento das despesas;

d) Gerência da documentação administrativa e contábil;

e) Representação do Fórum de Coordenadores, em atividades diversas; e

f) Suporte administrativo e de suprimentos.

A Secretaria Executiva tem sede em um núcleo escolhido pelo Fórum de Coordenadores.

Finalmente os núcleos e os grupos associados constituem a estrutura do Infohab. Os núcleos são vinculados a Universidades, centros de pesquisa e instituições sem fins lucrativos e sua criação é aprovada pelo Fórum de Coordenadores. Os grupos associados são unidades de captação de informação

com aprovação do Fórum de Coordenadores, avaliadas as necessidades do núcleo de origem e possibilidade de agregar resultados ao projeto.

Compete aos núcleos e grupos associados:

a) Busca, triagem, catalogação e classificação da informação;

b) Controle da qualidade de informação na área de domínio do Núcleo;

c) Arquivamento eletrônico;

d) Consultoria de catalogação na área de domínio do Núcleo;

e) Divulgação regional do Infohab;

f) Encaminhar solicitações às agências de fomento regionais;

g) Desenvolver iniciativas complementares à competência do Infohab com enfoques regionais em suas áreas de conhecimento específico, mediante aprovação do Fórum de Coordenadores.

São atribuídos a um núcleo, definido pelo Fórum de Coordenadores, a definição de parâmetros, normas e processos de tratamento da informação, bem como a montagem do sistema de informação. Na Figura 30 pode-se ter uma visão geral da estrutura organizacional do Infohab.

São duas as grandes atribuições do Infohab:

I. Manutenção de uma base de dados permanentemente e sistematicamente atualizada com as referências dos resultados de pesquisas, legislação federal, estadual e municipal, normas pertinentes, levantamentos governamentais e demais tipos de documentação disponível; e

II. Servir de repositório desta documentação em suas versões eletrônicas, respeitando-se os direitos autorais devidos.

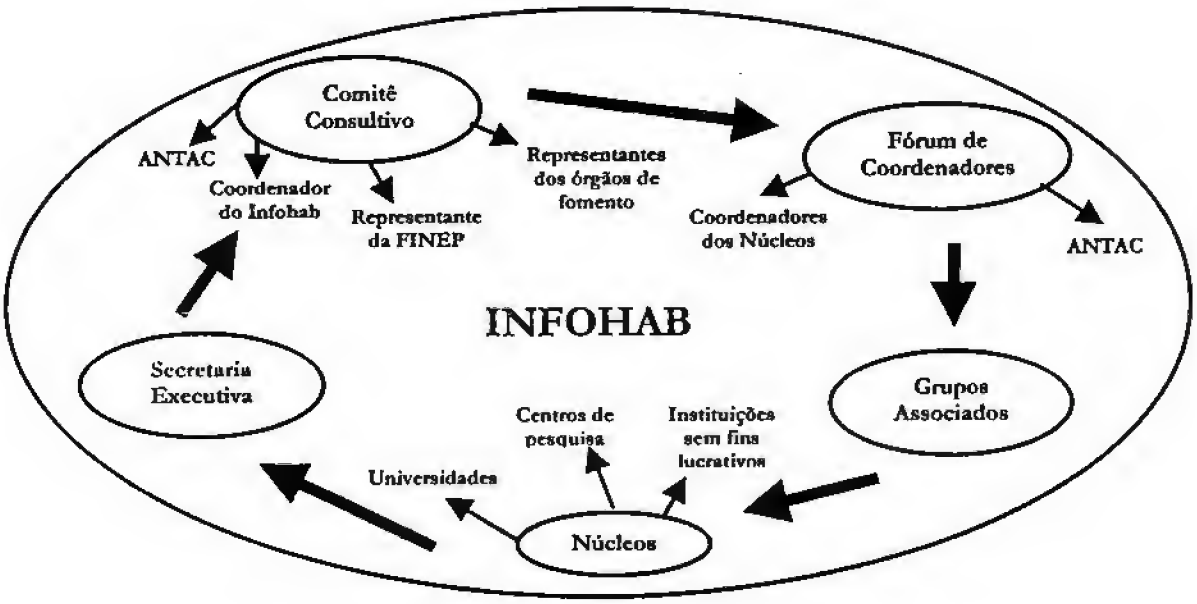


Figura 30 – Estrutura organizacional do Infohab

No início desta seção foi ressaltado que o universo possuía potencialidades que justificavam a sua significação para a esta pesquisa, sobretudo no seu caráter representativo, permanente e de abrangência nacional.

No tocante ao caráter representativo, o universo focalizado é uma das raras iniciativas de gestão do conhecimento da tecnologia do ambiente construído que se concretiza por meio da disponibilização de uma base de dados de alcance nacional voltada para especialistas, pesquisadores e interessados na questão da tecnologia do ambiente construído.

No que diz respeito ao caráter permanente, o Infohab vem incluindo no seu planejamento estratégico mudanças, a fim de dotar a atual estrutura organizacional do Centro de uma filosofia de auto-sustentação necessária para a sua consolidação e perenidade das atividades desenvolvidas pioneiramente hoje.

O caráter da abrangência nacional está também em grande medida no caráter representativo, já que se trata de uma iniciativa que envolve, a partir dos

seus núcleos, Universidades e Centros de Pesquisa espalhados pelas várias regiões do Brasil, garantindo com isso, credibilidade, representatividade e abrangência.

6.3- Caracterização da amostra selecionada

A amostra selecionada recaiu sobre o acervo de teses e dissertações inseridas no Infohab pela Centralizadora de Documentação e Informação (CEDIN), vinculada à Gerência Nacional de *Marketing* Interno. A amostra conta com 56 itens bibliográficos e a sua escolha deveu-se a três fatores:

- Todas as teses e dissertações que compõem a amostra são o resultado final de pesquisas realizadas pelo pessoal da CAIXA nas áreas de saneamento, desenvolvimento urbano e habitação, sendo portanto, uma amostra fiel dos assuntos de especialização do universo do trabalho: o ambiente construído. Correspondem também, à documentação do capital intelectual[11] de uma das mais importantes áreas-fim da CAIXA, o Desenvolvimento Urbano;

- A amostra corresponde ao terceiro tipo de bibliografia mais numeroso na base de dados do acervo CAIXA no Infohab, conforme a Tabela 6 a seguir; e

- Universo relativamente "fácil", em comparação a outros tipos de documentos disponíveis na organização;

[11] **Capital intelectual** – compreende o conhecimento que é de valor para uma organização construída de capital humano, capital estrutural e capital-cliente. Acredita-se que este fator possa ser analisado para permitir classificar a organização como rica ou pobre em informação (Halal & Kull, 2000 *apud* Tarapanoff, 2001).

Tabela 6 - Estatística da Base de dados do Infohab

TIPO DE ITEM BIBLIOGRÁFICO	QUANTIDADE
Anais de congresso	25
Artigo de congresso	15
Artigo de periódico	39
Dicionário	4
Documento sonoro	1
Especificação técnica	2
Folheto	528
Imagem em movimento	3
Legislação	3
Livro	795
Manual	42
Periódico	7
Tese e dissertação	56
Total de itens na base de dados	**1520**

Fonte: CEDIN/CAIXA

A CEDIN é responsável pela gestão da informação no âmbito da CAIXA, e foi por meio desta Unidade, que se estabeleceu a parceria entre o Infohab e a CAIXA como um Grupo Associado. O seu acervo é uma importante coleção sobre as áreas de saneamento, desenvolvimento urbano e habitação, tendo sido incorporado pela CAIXA quando da extinção do Banco Nacional de Habitação – BNH no ano de 1986. O BNH foi o responsável pelo desenvolvimento e implementação de toda a política de saneamento, desenvolvimento e habitação do Governo Federal, desde a década de 60, além de fomentador de pesquisas e estudos nestas áreas.

Existem inúmeras dificuldades na promoção e difusão de informação e de novas tecnologias no setor da construção, onde são maioria as pequenas e médias empresas, geograficamente dispersas e, em geral, com uma capacitação tecnológica relativamente simples. A demora na absorção das novidades e as dificuldades de interação entre as várias empresas são características de um setor altamente conservador, sem contar os obstáculos relativos ao acesso à informação com bases de dados dispersas e instrumentos de busca e recuperação da informação falhos.

Diante disto, a CAIXA, por intermédio da CEDIN, contribui para mudar este cenário, disponibilizando o seu acervo na área de desenvolvimento urbano no Infohab. Assim sendo, todo o acervo das áreas de saneamento, desenvolvimento urbano e habitação pertencente à CAIXA foi inserido na base de dados do Infohab, tornando-a uma referência de busca, recuperação e disseminação da informação na área de desenvolvimento urbano no Brasil. Fazem parte da estrutura organizacional da CEDIN duas funções técnicas de bibliotecários e um bibliotecário chefe, cujas atribuições incluem a seleção, indexação e inserção das publicações, além da manutenção da base de dados do Infohab como Grupo Associado previsto em Regimento.

6.4- Delineamento e histórico da pesquisa

6.4.1- Etapas da pesquisa

O desenvolvimento da metodologia está subdividido em etapas que englobaram procedimentos teóricos e práticos, inclusive com o desenvolvimento de um protótipo com aplicação da ferramenta de mineração de textos. Os procedimentos teóricos foram concentrados na construção de um referencial teórico constante da revisão de literatura que, durante a confecção do trabalho, subsidiou todas as decisões metodológicas ora apresentadas, além de fundamentar e contextualizar o problema. Os procedimentos de ordem prática seguiram a seqüência apresentada a seguir:

6.4.1.1- Estabelecimento da interface entre a pesquisa e o universo

Entre os meses de novembro e fevereiro dos anos de 2003 e 2004, foram estabelecidos diversos contatos com a CAIXA, por meio da CEDIN, a fim de apresentar a proposta de pesquisa para o estabelecimento de parceria, já que o *software* de mineração de textos para a construção do protótipo foi negociado pela equipe da Gerência Nacional de *Marketing* Interno da CAIXA em conjunto com a

Chefia da CEDIN com a Empresa Policentro fornecedora do software de mineração de textos, que disponibilizou a ferramenta **Br/*Search*** para utilização nesta pesquisa.

A necessidade de otimizar a performance da base de dados do Infohab, no que corresponde a melhoria dos procedimentos de indexação manual existentes, encaixou-se perfeitamente com a proposta da pesquisa, daí o interesse da CAIXA em negociar o *software* de mineração de textos, parte fundamental do trabalho, com a empresa fornecedora para a construção do protótipo.

Em abril de 2004, a Policentro, detentora dos direitos do *software* de mineração de textos **BR/*Search*** apresentou o produto em suas aplicações genéricas para a equipe da Gerência Nacional de *Marketing* Interno da CAIXA e a partir de então foram feitos diversos ajustes no protótipo desenvolvido na CAIXA, conforme iam avançando os entendimentos entre os atores envolvidos na pesquisa: a Policentro, a CEDIN e o autor desta pesquisa.

6.4.1.2- Definição da amostra

A partir dos fatores elencados na caracterização da amostra selecionada (ver página 130), ficou definido que "tese e dissertação" seria o tipo de item bibliográfico a ser considerado para a amostra e para construção do protótipo.

6.4.1.3- Extração da amostra do Infohab

A extração da amostra do Infohab se deu em dois momentos: **1º-** Os metadados de todas as teses e dissertações, bem como os mecanismos de recuperação por palavras-chave foram extraídos do universo do Infohab por meio de senha do administrador da Base na CEDIN, o que possibilitou a recuperação por palavras-chave em uma pesquisa, apenas das teses e dissertações que correspondem à amostra selecionada; e **2º-** Extração dos textos completos de cada metadado que compõe a amostra em formato *Adobe Portable Document*

Format – PDF do *software* Acrobat Reader 4.0 da empresa Adobe Systems Incorporated para um CD-Rom.

6.4.1.4- Construção do protótipo

A construção do protótipo iniciou-se com a separação dos metadados de cada item bibliográfico da amostra e depois com a submissão e captura dos 56 arquivos PDF, pelo *software* de mineração de textos BR/*Search*. Com a base pronta, passou-se a refinar os instrumentos de mineração de texto conforme a necessidade de utilização na pesquisa. Esta etapa correspondeu às partes 2- extração e 3- mineração propostos por Trybula (1999), conforme a Figura 29 da página 117, aonde incidiu a descoberta de conhecimento em textos por listas de conceitos-chave e por descrição de classes de textos descritos na seção de embasamento teórico da metodologia.

A extração e a mineração durante a construção do protótipo consumiu em torno de quatro meses para ser concluída, ou seja, em setembro de 2004 ficou pronta para testes. Isto se deveu a uma série de procedimentos de adaptação do banco de dados gerado para que o protótipo fosse o espelho da base de dados Infohab, condição básica para a realização dos testes de precisão. Simultaneamente ajustes nos textos gerados pelo *software* de mineração de textos se fizeram necessários, assim como na pesquisa e na geração de listas de conceitos chave que necessitou do apoio de um analista de sistema especialista em mineração de textos, pois o *software* oferecia soluções padronizadas e demasiado genéricas que necessitavam de adaptações.

6.4.1.5- O *software* de mineração de textos utilizado

O BR/*Search* foi desenvolvido em ferramenta OLAP[12] e é um produto voltado para o gerenciamento de informações não estruturadas. Trata-se de uma

[12] **Ferramenta OLAP** – programa que possibilita ao usuário a obtenção de informações armazenadas nas bases de dados dos data warehouses. Entre as suas principais funcionalidades estão o *drilling*, ou seja, detalhamento e o *slice & dice* ou seleção e visualização de porções de base de dados (Tarapanoff, 2001).

tecnologia de Gerenciamento Eletrônico de Documentos (GED) onde o tipo de dado "Texto Corrido" é tratado com o mesmo nível de importância que qualquer outro campo chave, tradicionalmente utilizado nos bancos de dados que suportam o modelo relacional, embora não siga tal modelo. Isto quer dizer, fundamentalmente, que palavras, expressões, etc., podem ser encontradas, mesmo que ela esteja explicitada apenas em um pequeno trecho de um texto. Esta potencialidade é característica dos *softwares* de mineração de textos.

O BR/*Search* pode ser considerado como um sistema de recuperação de informações baseada em um *software* de gerência, recuperação e mineração textual. Projetado para suportar grandes coleções de informações não estruturadas, possibilita que múltiplos usuários recuperem e analisem documentos armazenados praticamente em qualquer linguagem. Desta forma, uma das principais características do BR/*Search* é a identificação de padrões (*clusters*) a partir de texto completo. O usuário pode localizar e separar qualquer informação, em frações de segundo, com precisão, mesmo que esteja em uma única palavra ou frase em milhares de documentos. Basta que o usuário informe uma palavra ou frase.

6.4.1.6- A coleta de dados para os testes de precisão

Foram elaborados dois formulários de pesquisa (seção de anexos, páginas 211 e 213) que foram aplicados junto ao corpo técnico da CAIXA usuário do Infohab, a fim de realizar testes de precisão com a amostra selecionada e o protótipo. O formulário está anexado a este trabalho e foi estruturado para coletar dados relativos à pesquisa descrita em texto livre, palavras-chave tal como submetidas às duas bases de dados, inclusive com os operadores e campos pesquisados, o resultado das pesquisas e a posterior validação do usuário com a anotação de "útil" ou "inútil" para cada item bibliográfico recuperado. O formulário questionava ainda sobre dados do usuário tais como: nome, função unidade em

que trabalha, *e-mail* e telefone, com a finalidade de comunicar os resultados encontrados.

6.4.1.7- Seleção dos usuários para a coleta de dados

A seleção dos usuários seguiu o critério de gerência e consultoria técnica das áreas usuárias estratégicas do Infohab na Caixa, ou seja, Vice-Presidência de Desenvolvimento Urbano e Governo – VIURB/Diretoria de Parcerias e Apoio ao Desenvolvimento Urbano DIPUP e a Gerência Nacional de Normas e Padrões de Engenharia e Trabalho Social – GEPAD. Entre os 22 usuários consultados, 3 deles são bibliotecários da CEDIN que além de indexar a Base, são também usuários. Em anexo na página 222 segue a lista dos usuários .

A questão estratégica que incidiu na escolha dos usuários está no fato de que tais gerentes utilizam as informações do Infohab em decisões relativas aos programas de habitação prioritários da CAIXA, e no caso dos bibliotecários, a manutenção da base de dados em condições adequadas para a perfeita disseminação da informação.

6.4.1.8- Teste-piloto

Na primeira semana de dezembro de 2004 foi realizada uma coleta de dados a título de teste-piloto com uma Bibliotecária da CEDIN e um Consultor da GEPAD. O ensaio foi esclarecedor quanto à adaptação e compreensão dos respondentes acerca da expressão livre da pesquisa a ser realizada. Com esta experiência a técnica da entrevista de referência foi adotada nos procedimentos de pesquisa com o formulário definitivo.

6.4.1.9- Aperfeiçoamento do instrumento de coleta de dados

Como forma de otimizar os resultados e obter uma melhor visualização do cálculo do índice de precisão de cada pesquisa realizada, foi incluído no formulário um campo com a fórmula do cálculo da precisão. Esta adaptação visou também uma melhor organização dos resultados para os cálculos e inferências estatísticas que apontaram o ganho de precisão entre a lista de palavras-chave utilizadas na indexação manual e o protótipo.

6.4.1.10- Aplicação do instrumento de coleta de dados

Entre os dias 08 e 13 de dezembro de 2004 foi aplicado o instrumento de coleta de dados no Edifício da Matriz da CAIXA em Brasília, onde se situam a VIURB, DIPUP e a GEPAD com 19 gerentes e consultores. No dia 14 o instrumento foi aplicado com as 3 bibliotecárias da CEDIN também situada em Brasília. No total foram 22 entrevistas de referência com média de 10 minutos de duração cada.

6.4.1.11- Testes de precisão e validação dos usuários

O testes de precisão para a geração dos resultados de pesquisa foram realizados com os 44 formulários, 2 para cada entrevistado, no dia 15 de dezembro de 2004. Neste teste as mesmas palavras-chave com os respectivos operadores no campo palavra-chave foram aplicadas de forma idêntica ao protótipo e à lista de palavras-chave utilizadas na indexação manual.

Entre os dias 16 e 20 de dezembro o resultado das pesquisas nas bases foi encaminhado para a apreciação e validação dos usuários com a seguinte anotação para cada item bibliográfico recuperado: documento útil (U) ou documento inútil (I). Os resultados foram apresentados em duas listas distintas, uma com o resultado da pesquisa na base do Infohab e a outra lista com o

resultado da pesquisa no protótipo. Este procedimento visou garantir a correta aplicação da fórmula do cálculo do índice de precisão para cada uma das bases.

Na tabela 7 é apresentada uma visão geral da correlação entre pressupostos, variáveis e ação metodológica propostos para a pesquisa, a fim de complementar a elucidação dos métodos, técnicas e instrumentos utilizados:

Tabela 7 - Pressupostos, variáveis e ação metodológica da pesquisa

Pressupostos	Variáveis	Ação Metodológica
1° Pressuposto: Independentemente da utilização da ferramenta de mineração de textos ou da lista de palavras-chave utilizadas na indexação manual, não há ganho significativo no índice de precisão quando do processo de busca e recuperação da informação.	**Variáveis: A)** Índice médio percentual de precisão calculado a partir da média entre os índices resultantes da lista de palavras-chave utilizadas na indexação manual; **B)** Índice médio percentual de precisão calculado a partir da média entre os índices resultantes do uso da ferramenta de mineração de textos; **C)** Comparação entre os índices médios percentuais de precisão com base nos resultados obtidos da lista de palavras-chave utilizadas na indexação manual e o uso da ferramenta de mineração de textos.	- Cálculo do índice de precisão resultante da lista de palavras-chave utilizadas na indexação manual e do protótipo com uso da ferramenta de mineração de textos por meio da fórmula da precisão; - Cálculo dos índices médios de percentuais de precisão obtidos da lista de palavras-chave utilizadas na indexação manual e do protótipo com uso da ferramenta de mineração de textos e a comparação dos resultados apurados; - Comparação dos resultados dos índices médios percentuais de precisão apurados.
2° Pressuposto: O uso da ferramenta de mineração de textos na busca e recuperação da informação trará sempre como resposta uma maior quantidade de itens bibliográficos do que a lista de palavras-chave utilizadas na indexação manual.	**Variáveis: A)** Cálculo do número total de itens bibliográficos recuperados da lista de palavras-chave utilizadas na indexação manual; **B)** Cálculo do número total de itens bibliográficos recuperados com uso da ferramenta de mineração de textos; **C)** Comparação dos números totais de itens bibliográficos recuperados da lista de palavras-chave utilizadas na indexação manual e do uso da ferramenta de mineração de textos.	- Cálculo da quantidade de itens bibliográficos recuperados na lista de palavras-chave utilizadas na indexação manual; - Cálculo da quantidade de itens bibliográficos recuperados no protótipo com uso da ferramenta de mineração de textos; - Comparação dos resultados das quantidades de itens bibliográficos apurados.
3° Pressuposto: Quando termos específicos da Base de dados do Infohab são submetidos à lista de palavras-chave utilizadas na indexação manual na busca e recuperação da informação, nem sempre são encontrados itens bibliográficos, enquanto que os mesmos termos quando submetidos ao Protótipo com aplicação da mineração de textos, sempre irão recuperar algum item bibliográfico.	**Variáveis: A)** Resultados percentuais nulos obtidos na contagem de itens bibliográficos recuperados na lista de palavras-chave utilizadas na indexação manual; **B)** Resultados percentuais nulos obtidos na contagem de itens bibliográficos recuperados com a utilização da ferramenta de mineração de textos.	- Separação, contagem e comparação dos resultados nulos obtidos da lista de palavras-chave utilizadas na indexação manual e do protótipo com uso da ferramenta de mineração de textos.

6.4.1.12- Tratamento dos dados

O tratamento dos dados deu-se em duas etapas:

I) Cálculo do índice de precisão: com os resultados validados por 22 usuários dos itens bibliográficos recuperados da lista de palavras-chave utilizadas na indexação manual e do uso da ferramenta de mineração de textos, foi realizado o cálculo do índice de precisão. Para tanto a fórmula usada foi:

$$\textit{Precisão} = \frac{\textit{Número de } \textbf{\textit{documentos úteis}} \textit{ recuperados pelo sistema}}{\textit{Número } \textbf{\textit{total de documentos encontrados}} \textit{ pelo sistema}} \; X\,100$$

II) Tabulação dos dados: com os resultados dos índices de precisão apurados, foi possível calcular e inferir se no estudo de caso do Infohab, a lista de palavras-chave utilizadas na indexação manual trouxe um índice de precisão maior na busca e recuperação da informação do que com o uso da ferramenta de mineração de textos, além das inferências que complementam a tabulação dos dados na comprovação ou refutação dos pressupostos da pesquisa.

7- ANÁLISE DOS DADOS E COMPROVAÇÃO DOS PRESSUPOSTOS

Nesta seção apresentaremos os resultados por meio da análise dos dados coletados. Faremos isto a partir da análise dos dados sob aspectos descritivos, ou seja, de posse das variáveis da pesquisa, descreveremos os resultados alcançados para a comprovação dos pressupostos e discussão das teses.

7.1- Cálculo do índice de precisão

O cálculo do índice de precisão obtido com a Base de dados do Infohab foi realizado a partir de 22 testes aplicados com 22 especialistas da CAIXA. Para cada teste foi calculado o índice de precisão dado pelo número de documentos úteis recuperados pelo sistema sobre o número total de documentos encontrados pelos sistema. O resultado é obtido em termos percentuais, já que se multiplica a proporção por 100. Assim, cada um dos usuários contribui, julgando se cada item bibliográfico recuperado pela Base de dados do Infohab é útil ou inútil. Esta medida é o elemento essencial para o cálculo do índice de precisão.

Na fórmula utilizada para o cálculo do índice de precisão, pode-se ilustrar o papel da validação (julgamento) do usuário quanto a utilidade ou não de um item bibliográfico recuperado:

$$\textit{Precisão} = \frac{\textit{Número de documentos úteis recuperados pelo sistema}}{\textit{Número total de documentos encontrados pelo sistema}} X\ 100 = \frac{a}{a + c}$$

Onde: *a* = documentos úteis e recuperados e *c* = documentos inúteis e recuperados.

O julgamento do usuário incide tanto em "*a*" quanto em "*c*", ou seja, o cálculo do índice de precisão só pode ser realizado com a validação do usuário. Daí a importância da entrevista de referência realizada com cada um dos 22 usuários para a metodologia adotada na realização da pesquisa.

Outro fator preponderante é que, para a realização dos julgamentos, os usuários necessitam ser especialistas no assunto (tema) dos possíveis itens bibliográficos a serem recuperados. No caso em questão, todos os usuários são especialistas no assunto ambiente construído, tema da Base de dados do Infohab.

A Tabela 8 a seguir apresenta os 22 testes de precisão realizados na Base de dados do Infohab com os respectivos resultados dos cálculos da precisão:

Tabela 8 - Testes de precisão realizados na Base de dados do Infohab

Base de dados do Infohab aplicada aos testes	Resultados dos cálculos dos índices de precisão (%)
Teste 1 (T1)	0,0%
Teste 2 (T2)	16,66%
Teste 3 (T3)	50%
Teste 4 (T4)	100%
Teste 5 (T5)	0,0%
Teste 6 (T6)	20%
Teste 7 (T7)	0,0%
Teste 8 (T8)	50%
Teste 9 (T9)	50%
Teste 10 (T10)	0,0%
Teste 11 (T11)	60%
Teste 12 (T12)	15,38%
Teste 13 (T13)	0,0%
Teste 14 (T14)	50%
Teste 15 (T15)	33,33%
Teste 16 (T16)	100%
Teste 17 (T17)	50%
Teste 18 (T18)	37,5%
Teste 19 (T19)	42,85%
Teste 20 (T20)	20%
Teste 21 (T21)	57,14%
Teste 22 (T22)	0,0%

A análise do Gráfico 1, relativo ao índice de precisão obtido com a Base de dados do Infohab, mostra uma dispersão[13] quanto aos resultados dos índices de

[13] **Dispersão** – Flutuação de uma variável em um grupo de observações.

precisão, apontando para uma regularidade apenas relativa no nível intermediário do Gráfico.

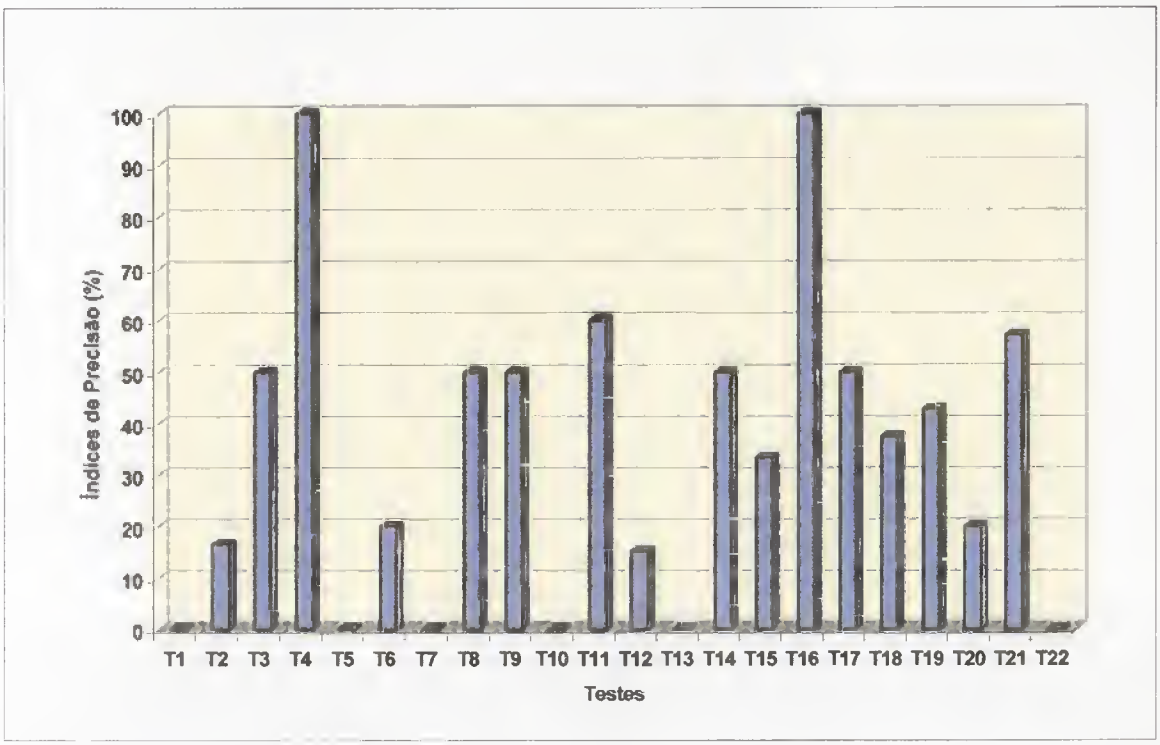


Gráfico 1 - Índice de precisão obtido com a Base de dados do Infohab

Outra constatação a que pode-se chegar ao analisar o Gráfico 1, diz respeito à quantidade de testes com resultado nulo. Este fato ocorre em seis situações. O índice máximo de 100% de precisão foi alcançado em dois testes (4 e 16 respectivamente). Aliás, cabe ressaltar que, além destes dois casos em que o índice de precisão alcançou o nível máximo, as melhores performances apareceram em índices bem inferiores, ou seja, em apenas dois momentos. Um dos casos com índice de 60% (teste 11) e em outro com 57,14% (teste 21) na base de dados do Infohab.

Os resultados mais baixos quanto ao índice de precisão, além dos seis testes com resultado nulo (testes 1, 5, 7, 10, 13 e 22), foram detectados com os

testes de número 2 com 16,66% de precisão, com o teste 6 com 20%, teste 12 com 15,38% e teste 20 com o índice de precisão de 20%.

Estes resultados confirmam a grande irregularidade dos retângulos do Gráfico 1, além do grande número de índices de precisão ocorrendo na parte intermediária do Gráfico, entre 30 e 50%, precisamente em oito testes.

Este panorama deve rebaixar a média do índice de precisão obtido com a Base de dados do Infohab que será apresentada mais à frente.

Nos resultados dos testes de precisão obtidos com a Base de Dados do Infohab, em dois deles que perfazem 9,09% do total de 22 testes, o índice de precisão foi de 100%; em 7 testes ou 31,81% do total de 22 testes, o índice de precisão variou entre 60 e 50%; em outros 7 testes, 31,81% do total dos testes, o índice de precisão variou entre 42,85 e 15,38% e finalmente em 6 testes ou 27,27% do total de 22 testes, o índice de precisão foi de 0,0%.

Estes dados podem ser melhor visualizados na Tabela 9:

Tabela 9 – Percentual do total de testes e índice de precisão na Base de dados do Infohab

Base de dados do Infohab aplicada aos testes	Percentual do total de 22 testes	Índice de precisão
Dois testes (4 e 16)	9,09%	100%
Sete testes (3, 8 ,9 ,11, 14, 17 e 21)	31,81%	De 50 a 60%
Sete testes (2, 6, 12, 15, 18, 19 e 20)	31,81%	De 42,85 a 15,38%
Seis testes (1, 5, 7, 10, 13 e 22)	27,27%	0%

No Gráfico 2, a seguir, fica melhor representada a variação percentual[14] dos testes de precisão, mostrando claramente que os índices com 0% de precisão são a segunda maior concentração de testes, ou seja, 27,27% (6 testes) do total. Este

[14] **Variação percentual** – conceito estatístico fundamental na pesquisa científica. Virtualmente nenhum conhecimento científico seria possível se o fenômeno não variasse (Kerlinger, 1980). Variação medida em termos percentuais.

resultado é o mais significativo no rebaixamento do valor médio[15] de precisão obtido com a Base de dados do Infohab.

Na segunda maior concentração de testes, tem-se dois grupos com exatos 31,81% (7 testes) do total de 22 realizados com variação percentual de 50 a 60% e 42,85 a 15,38% respectivamente. Nos dois testes restantes, ou seja, em 9,09% dos casos foi alcançado 100% de precisão.

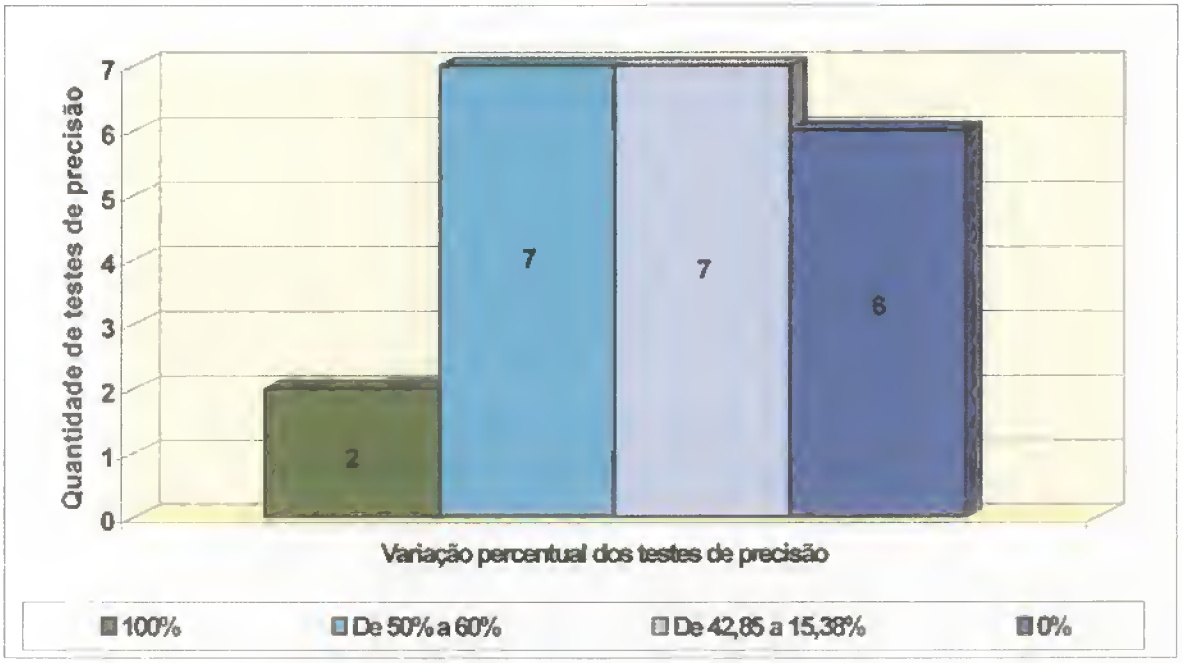


Gráfico 2 – Resultado dos testes de precisão realizados na Base de Dados do Infohab

A fim de complementar as análises sobre o significado de cada conjunto de testes em relação a variação percentual do índice de precisão, bem como a representatividade destes em relação a sua participação percentual no universo dos 22 testes obtidos com a Base de dados do Infohab, é apresentado o Gráfico 3, onde está disposta nos eixos "x" e "y" a representatividade percentual dos grupos de testes com as suas respectivas quantidades.

O Gráfico confirma a análise feita sobre o rebaixamento do valor médio de precisão obtido com a Base de dados do Infohab no grupo que responde com uma

[15] **Valor médio** – valor obtido a partir do cálculo da média de um dado conjunto de valores.

variação no índice de precisão entre 42,85 e 15,38%, além da grande quantidade de índices nulos.

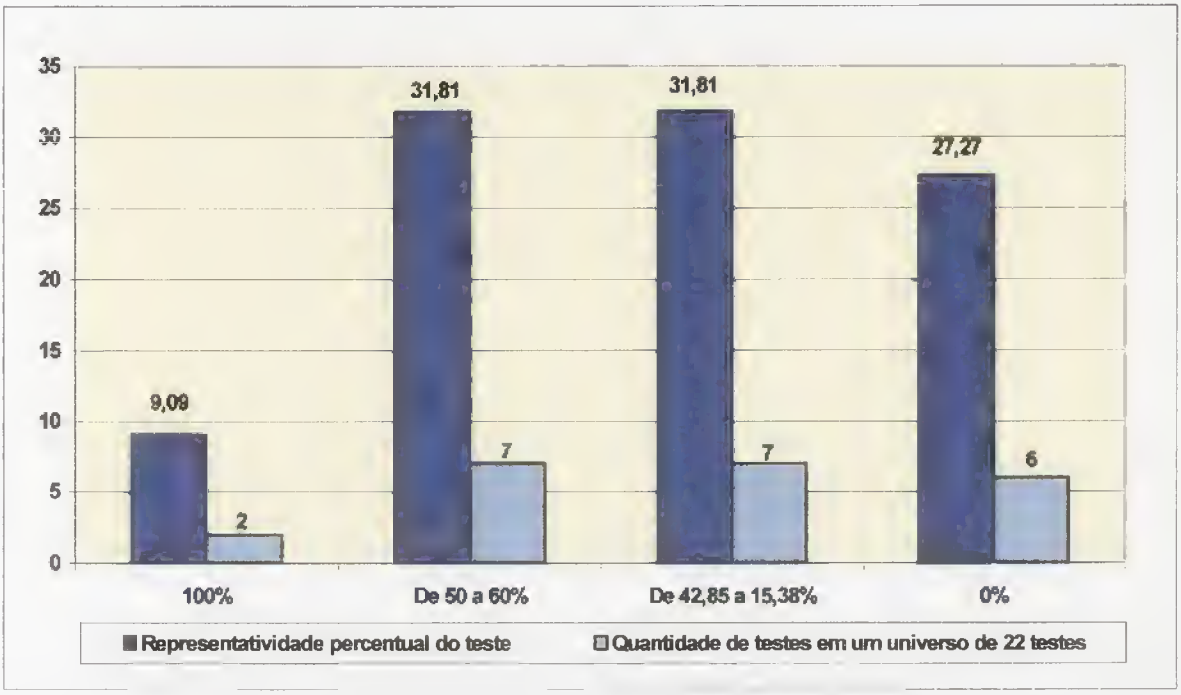


Gráfico 3 - Representatividade dos testes de precisão obtidos com a Base de dados do Infohab

Na Tabela 10, a seguir, são apresentados os 22 testes de precisão realizados no Protótipo com aplicação da mineração de textos com os respectivos resultados dos cálculos da precisão:

Tabela 10 - Testes de precisão realizados no Protótipo com aplicação da mineração de textos

Protótipo aplicado aos testes	Resultados dos cálculos dos índices de precisão (%)
Teste 1 (T1)	18,18%
Teste 2 (T2)	100%
Teste 3 (T3)	36,66%
Teste 4 (T4)	30,76%
Teste 5 (T5)	0,0%
Teste 6 (T6)	38,88%
Teste 7 (T7)	18,75%
Teste 8 (T8)	70%
Teste 9 (T9)	75%

Teste 10 (T10)	36,84%
Teste 11 (T11)	21,42%
Teste 12 (T12)	44,44%
Teste 13 (T13)	33,33%
Teste 14 (T14)	21,73%
Teste 15 (T15)	10,52%
Teste 16 (T16)	34,61%
Teste 17 (T17)	17,85%
Teste 18 (T18)	11,11%
Teste 19 (T19)	23,8%
Teste 20 (T20)	16,66%
Teste 21 (T21)	51,16%
Teste 22 (T22)	72,41%

A análise do Gráfico 4, relativa ao índice de precisão obtido com o Protótipo com aplicação da mineração de textos, complementa a apresentação dos resultados, mostrando também dispersão quanto aos resultados dos índices de precisão, apontando para uma regularidade relativa no nível situado entre 40 e 20% (10 testes) de precisão, abaixo do nível intermediário do Gráfico.

Outra análise que pode ser feita ao verificar o Gráfico 4, diz respeito ao fato de apenas um teste de precisão ter resultado nulo. Coincidentemente o índice máximo de 100% de precisão foi alcançado também em um teste de precisão (T2). Além deste caso em que o índice de precisão alcançou o nível máximo, as melhores performances apareceram em índices entre 75 e 70%, em três testes: um caso com índice de 75% (teste 9), outro com 72,41% (teste 22) e o teste 8 com 70% de índice de precisão obtido no Protótipo com uso da mineração de textos.

Os resultados mais baixos quanto ao índice de precisão, além do teste 5 com resultado nulo, foram os detectados com os testes de número 15, 17 e 18 com 10,52%, 17,85% e 11,11% respectivamente.

Estes resultados além de confirmarem a grande irregularidade dos retângulos do Gráfico 4, assim como ocorreu com testes de precisão na Base de dados do Infohab, também rebaixam a média do índice de precisão obtida no

Protótipo, com aplicação da mineração de textos que será apresentada nas análises subseqüentes.

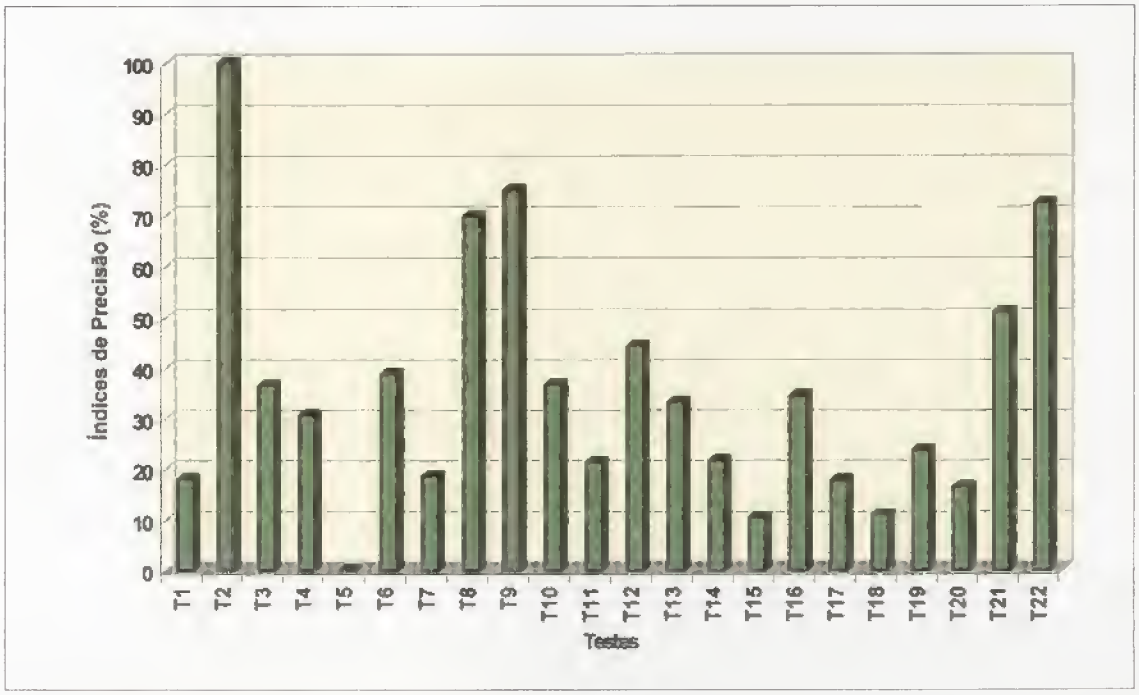


Gráfico 4 – Índice de precisão obtido com o Protótipo

Nos resultados dos 22 testes de precisão obtidos com o Protótipo com aplicação da mineração de textos, em um deles que perfaz 4,54% do total de 22 testes, o índice de precisão foi de 100%; em 3 testes ou 13,63% do total de 22 testes o índice de precisão variou entre 75 e 70%; em outros 8 testes que representam 36,36% do total, o índice de precisão variou entre 51,16 e 30,76%; em 9 testes, 40,9% do total, o índice de precisão variou entre 23,8 e 10,52% e finalmente em apenas um teste ou 4,54% do total de 22 testes, o índice de precisão foi de 0,0%.

Estes dados podem ser melhor visualizados na Tabela 11:

Tabela 11 – Percentual do total de testes e índice de precisão no Protótipo com aplicação da mineração de textos

Base de dados do Infohab aplicada aos testes	Percentual do total de 22 testes	Índice de precisão
Um teste (2)	4,54%	100%
Três testes (8, 9 e 22)	13,63%	De 75 a 70%
Oito testes (3, 4, 6, 10, 12, 13, 16 e 21)	36,36%	De 51,16 a 30,76%
Nove testes (1, 7, 11, 14, 15, 17, 18, 19 e 20)	40,9%	De 23,8 a 10,52%
Um teste (5)	4,54%	0%

No Gráfico 5, a seguir, fica melhor representada a variação percentual dos testes de precisão, mostrando claramente que o índice com 0% é a exceção dentre os testes, com 4,54% do total de 22. Na outra ponta, a exceção fica com o índice de 100%, também representando 4,54% do total. A maior concentração de testes encontra-se no grupo com 40,9% (9 testes) do total de 22 realizados com variação percentual de 23,8 a 10,52%.

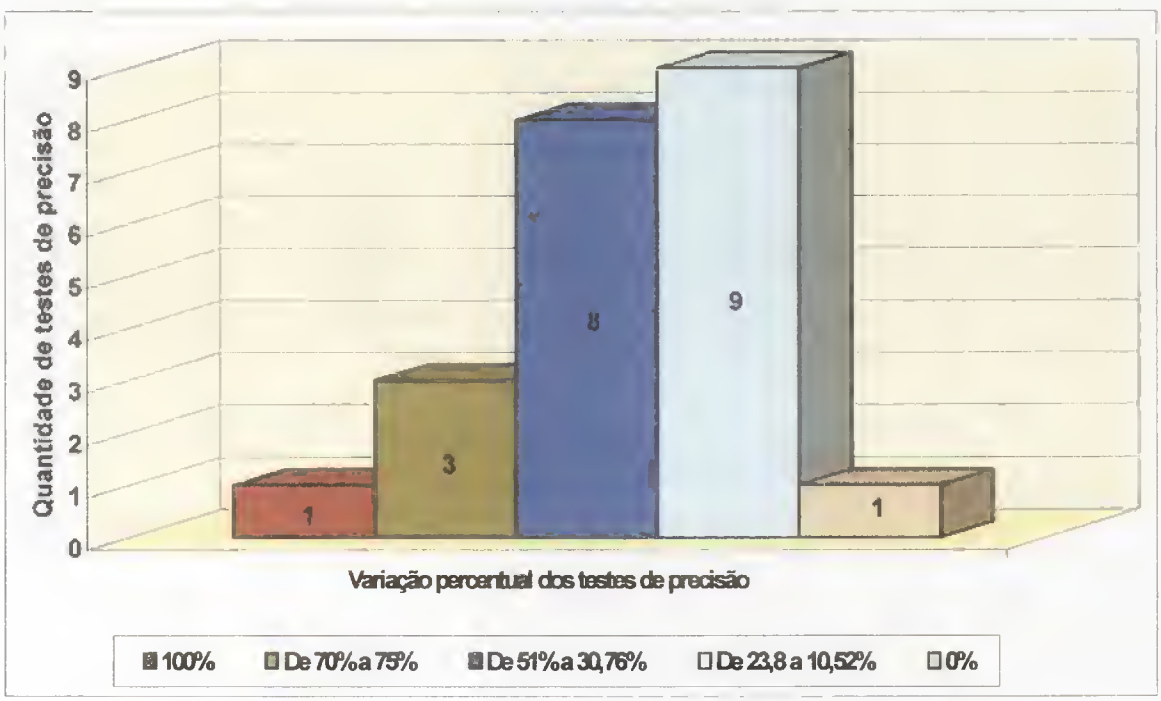


Gráfico 5 – Resultado dos testes de precisão realizados no protótipo

Este resultado é o mais significativo no rebaixamento do valor médio de precisão obtido com a Base de dados do Infohab. Os outros grupos são 3 testes com variação entre 75 e 70% e oito testes com variação entre 51,16 e 30,76%.

A fim de complementar as análises sobre o significado de cada conjunto de testes em relação à variação percentual do índice de precisão, bem como a representatividade destes em relação a sua participação percentuai no universo dos 22 testes obtidos com o Protótipo com aplicação da mineração de textos, é apresentado o Gráfico 6, onde está disposta nos eixos "x" e "y" a representatividade percentual dos grupos de testes com as suas respectivas quantidades.

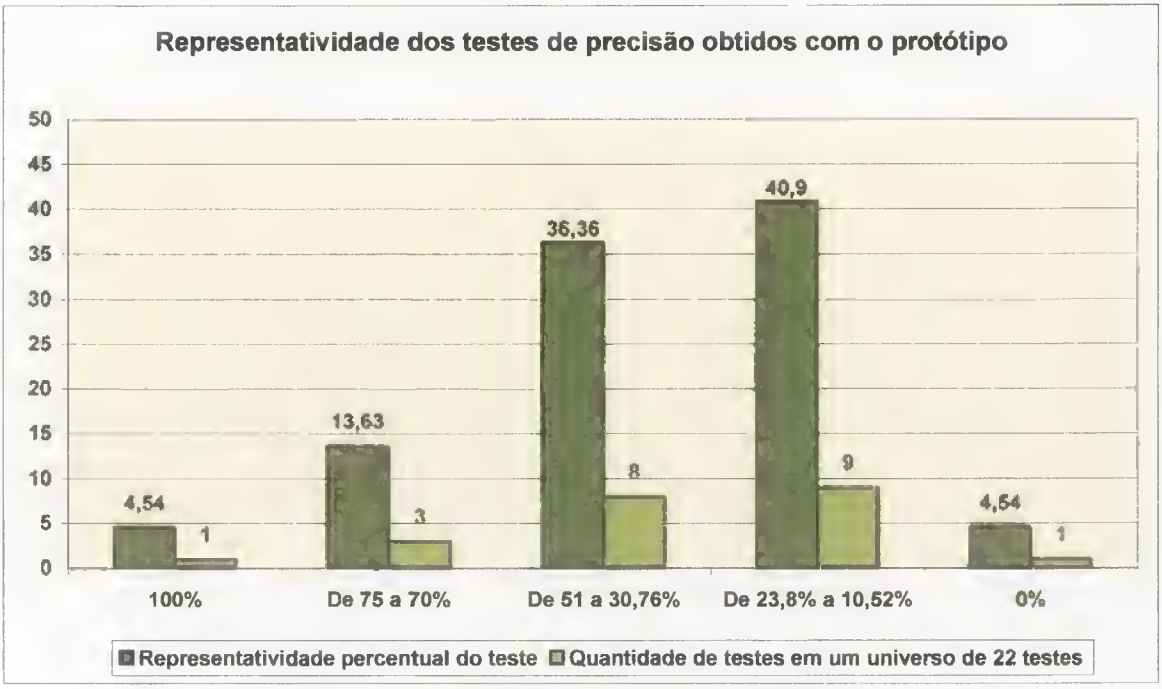


Gráfico 6 - Representatividade dos testes de precisão obtidos com o protótipo

O Gráfico confirma, mais uma vez, a análise feita sobre o rebaixamento do valor médio de precisão obtido com o Protótipo no grupo que responde com uma variação no índice de precisão entre 23,8 e 10,52% (9 testes).

Também chama a atenção, diferentemente do gráfico 3 (representatividade dos testes de precisão obtidos com a Base de dados do Infohab), o resultado do índice de precisão de 100% em uma ponta do gráfico e o resultado nulo na outra ponta. Provavelmente, seja esta a maior diferença obtida com testes de precisão entre a Base de dados do Infohab e o Protótipo com aplicação da mineração de textos.

Na Tabela 12 é apresentada a comparação entre os resultados obtidos nos índices de precisão na Base de dados do Infohab e o Protótipo com aplicação da mineração de textos, para que se possa conferir o ganho de precisão entre ambos.

Tabela 12 – Comparativo dos resultados dos testes de precisão realizados com a Base de dados do Infohab e com o Protótipo com mineração de textos

Base de dados do Infohab e o Protótipo com mineração de textos aplicados aos testes	Resultados com a Base do Infohab (Índice de Precisão)	Resultados com o Protótipo (Índice de Precisão)
Teste 1 (T1)	0,0%	18,18%
Teste 2 (T2)	16,66%	100%
Teste 3 (T3)	50%	36,66%
Teste 4 (T4)	100%	30,76%
Teste 5 (T5)	0,0%	0,0%
Teste 6 (T6)	20%	38,88%
Teste 7 (T7)	0,0%	18,75%
Teste 8 (T8)	50%	70%
Teste 9 (T9)	50%	75%
Teste 10 (T10)	0,0%	36,84%
Teste 11 (T11)	60%	21,42%
Teste 12 (T12)	15,38%	44,44%
Teste 13 (T13)	0,0%	33,33%
Teste 14 (T14)	50%	21,73%
Teste 15 (T15)	33,33%	10,52%
Teste 16 (T16)	100%	34,61%
Teste 17 (T17)	50%	17,85%
Teste 18 (T18)	37,5%	11,11%
Teste 19 (T19)	42,85%	23,8%
Teste 20 (T20)	20%	16,66%
Teste 21 (T21)	57,14%	51,16%
Teste 22 (T22)	0,0%	72,41%

A análise do Gráfico 7 relativo à comparação entre os índices de precisão da Base de dados do Infohab e do Protótipo com mineração de textos, permite

inferir que a maioria dos testes 24 ou 54,54% do total de 44 testes, obteve entre 0 e 35% de índice de precisão, sendo que só de índices de precisão nulos a porcentagem foi de 15,9% (7 testes) do total de 44. Este dado permite confirmar o que já havia sido constatado quando da análise em separado dos índices de precisão da Base de dados e do Protótipo, ou seja, o rebaixamento da média do índice de precisão em ambos os casos.

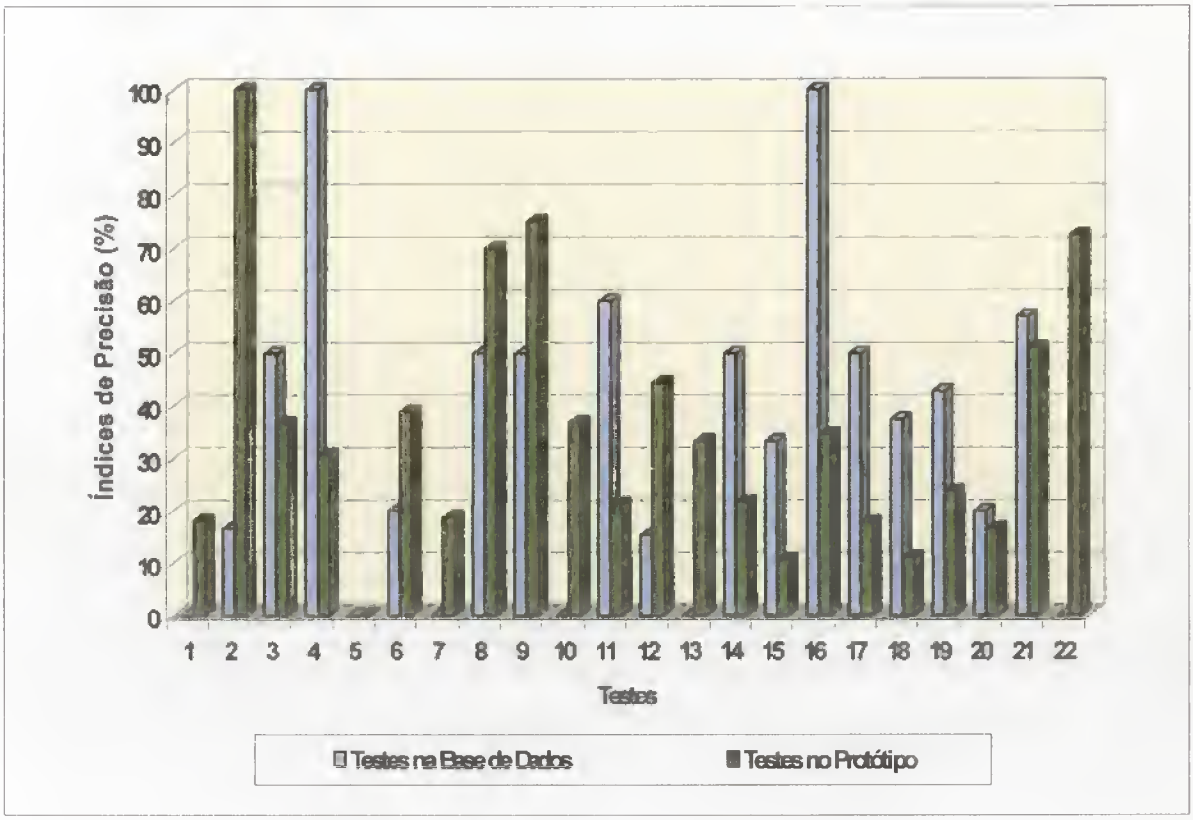


Gráfico 7 – Comparação entre os índices de precisão da Base de dados do Infohab e do Protótipo

Quando se analisa a melhor performance dos índices de precisão, conclui-se que em ambos os casos o índice atingiu 100% em apenas três momentos, o que em termos percentuais representa apenas 6,81% do total de 44 testes. A exigüidade de boas performances também se confirma na faixa de índices de

precisão entre 36 e 59%, foram treze testes ou 29,54% do total e na faixa compreendida entre 60 e 80% de índices de precisão com quatro testes ou 9,09% do total.

Estes resultados confirmam mais uma vez a grande irregularidade na variação das barras do histograma[16] representadas no Gráfico 7, a exemplo do que ocorre com a linha seqüencial irregular do Gráfico 1 com o índice de precisão obtido com a Base de dados e 4 com o índice de precisão obtido com o Protótipo. Isto demonstra a baixa média do índice de precisão obtido com a Base de dados e com o Protótipo, além de apontar de forma preliminar, para um ganho de precisão quase igual, utilizando a Base de dados do Infohab ou o Protótipo com o emprego da mineração de textos.

7.1.1- Comprovação do 1° pressuposto

1° Pressuposto: Independentemente da utilização da ferramenta de mineração de textos ou da lista de palavras-chave utilizadas na indexação manual, não há ganho significativo no índice de precisão quando do processo de busca e recuperação da informação.

Nos resultados dos testes de precisão (22 testes) obtidos com a Base de Dados do Infohab, o índice médio de precisão, calculado a partir da média entre os 22 índices resultantes, foi de 34,22%.

No cálculo dos índices de precisão realizados, esta tendência já se confirmava com baixa média, que concentrou 13 testes na faixa entre 42,85 a 0% de variação no índice de precisão. Estes testes corresponderam a 59,09% do total de testes. Desta forma, o índice médio de precisão atingiu 34,22% e pode ser examinado no Gráfico 8 a seguir:

[16] **Histograma** – gráfico estatístico que consiste em retângulos contíguos com base nas faixas de valores da variável e com área igual à freqüência relativa da respectiva faixa (Magalhães & Lima, 2001).

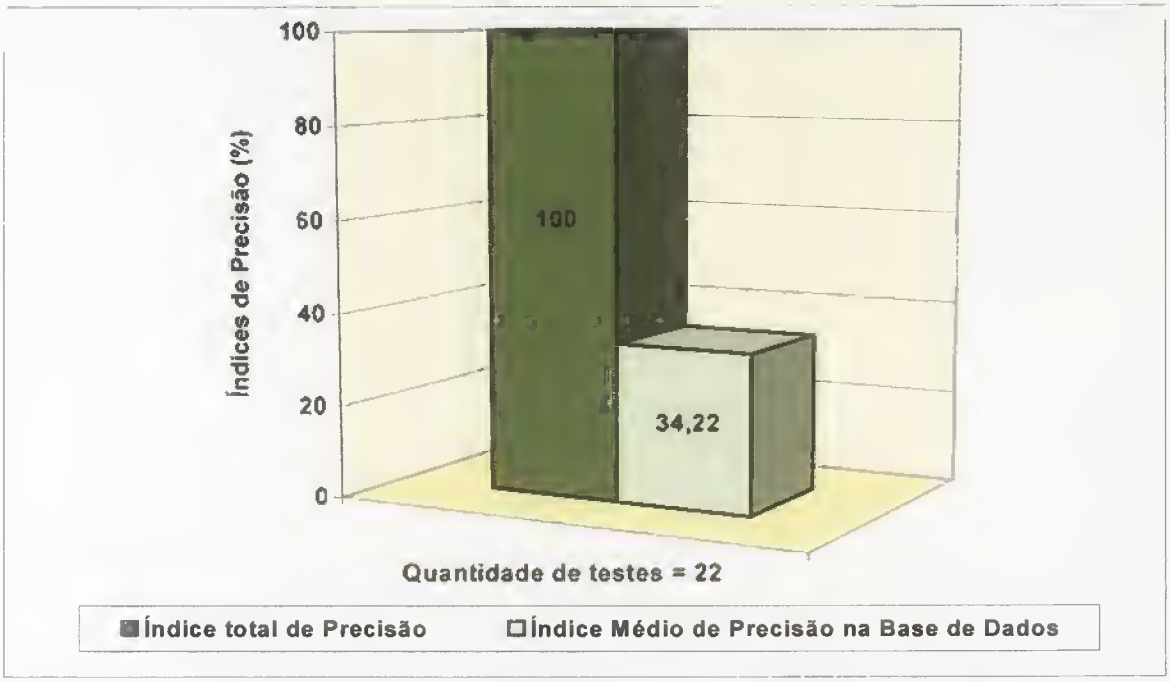


Gráfico 8 - Índice médio de precisão obtido na Base de Dados do Infohab

Nos resultados dos testes de precisão (22 testes) obtidos com o Protótipo com mineração de textos, o índice médio de precisão, calculado a partir da média entre os 22 índices resultantes, foi de 35,64%.

Também no caso do Protótipo, esta tendência de um baixo índice de precisão já se desenhava. Nos cálculos dos índices de precisão 9 testes ou 49,9% do total de 22, apresentaram variação percentual entre 23,8 e 10,52%, ou seja, um índice baixo e parecido com os resultados obtidos com a Base de dados do Infohab. Assim sendo, o índice médio de precisão atingiu 35,64% e pode ser analisado no Gráfico 9 a seguir:

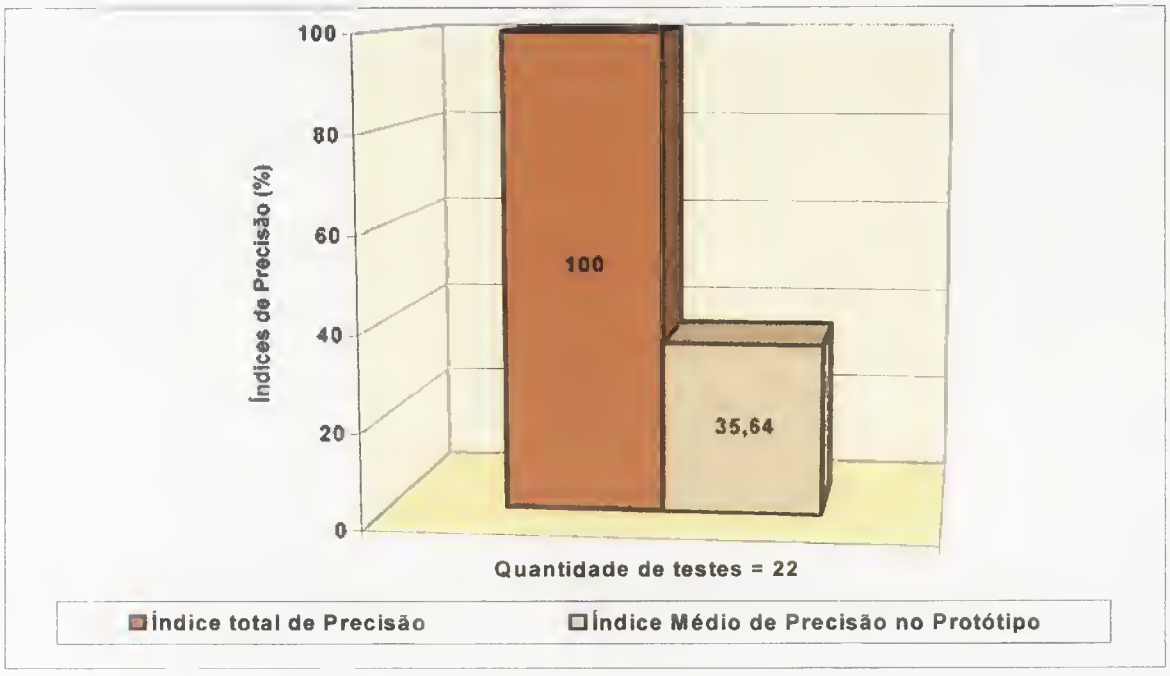


Gráfico 9 - Índice médio de precisão obtido no Protótipo

A comparação dos índices médios de precisão obtidos com a Base de dados do Infohab e com o Protótipo, com uso da ferramenta de mineração de textos, comprova o baixo índice na média de precisão obtido. Além disto, pela análise do Gráfico 10, os índices médios de precisão são muito parecidos com uma diferença de apenas 1,4 ponto percentual a mais para os resultados obtidos no Protótipo com uso da ferramenta de mineração de textos.

Cabe ressaltar que a diferença entre o maior índice médio de precisão obtido com o Protótipo, em relação ao índice máximo que seria de 100%, é de 64,36%. Portanto, uma diferença considerável para a análise empreendida.

Finalmente, o fator que mais chama a atenção e é decisivo para complementar a análise e comprovar o 1º pressuposto da pesquisa, está no fato de que não há como afirmar categoricamente que o maior índice médio de precisão que foi obtido com o Protótipo, traz ganho significativo no processo de busca e recuperação da informação. A diferença de 1,4% é insignificante para comprovar ganho significativo.

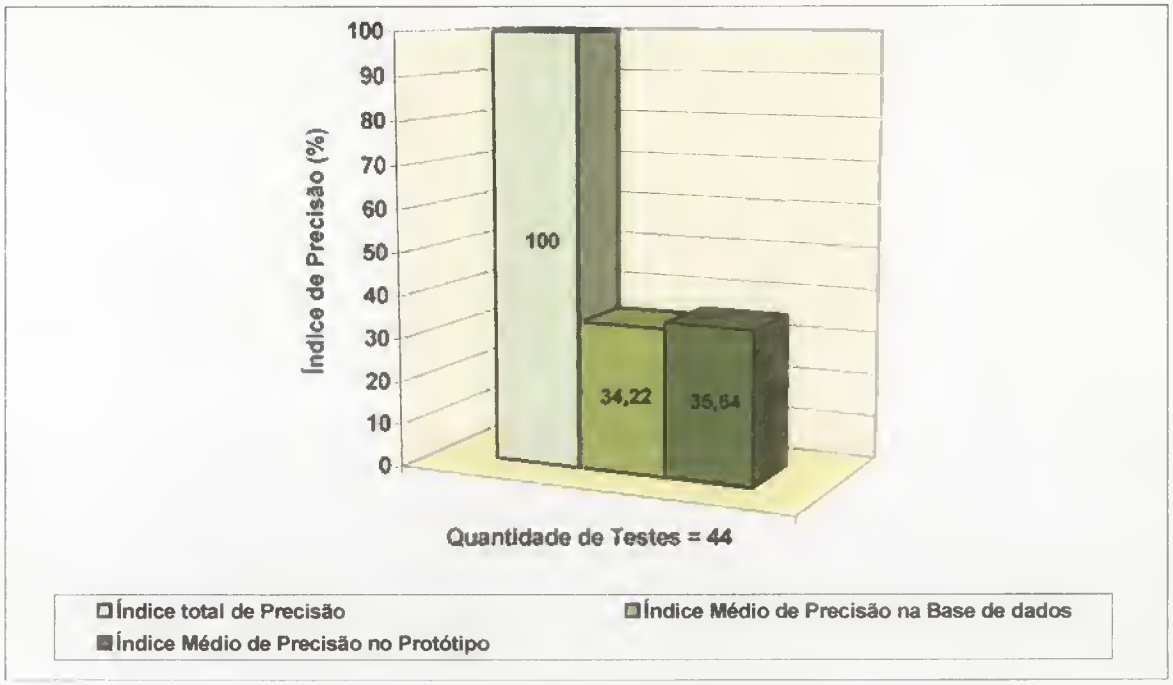


Gráfico 10 - Comparação entre os índices médios de precisão da Base de dados do Infohab e do Protótipo

Para confirmar a análise feita até aqui, foram ainda contabilizados os resultados dos índices percentuais de precisão de cada um dos 44 testes na Base de dados do Infohab e no Protótipo com uso da ferramenta de mineração de textos.

50% dos testes (11 testes) na Base de Dados do Infohab resultaram em um índice de precisão maior do aquele encontrado com os testes realizados no Protótipo.

Em um teste de precisão o resultado foi nulo tanto na Base de Dados do Infohab como no Protótipo.

A Base de dados apresenta um índice de precisão geral maior que o do protótipo, entretanto a diferença entre as duas contagens é pequena, mais ainda pelo fato de que em dois testes o resuitado foi nuio tanto para a Base de dados quanto para o Protótipo. O Gráfico 11 resume os resultados descritos:

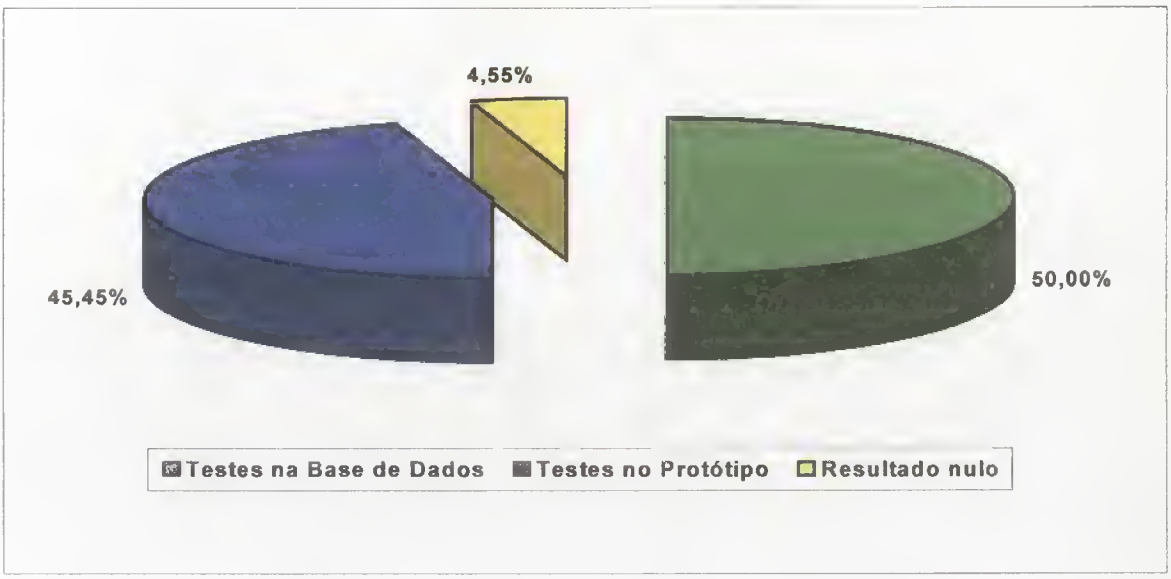


Gráfico 11 - Comparação dos valores totais dos cálculos de precisão entre a Base de dados do Infohab e o Protótipo

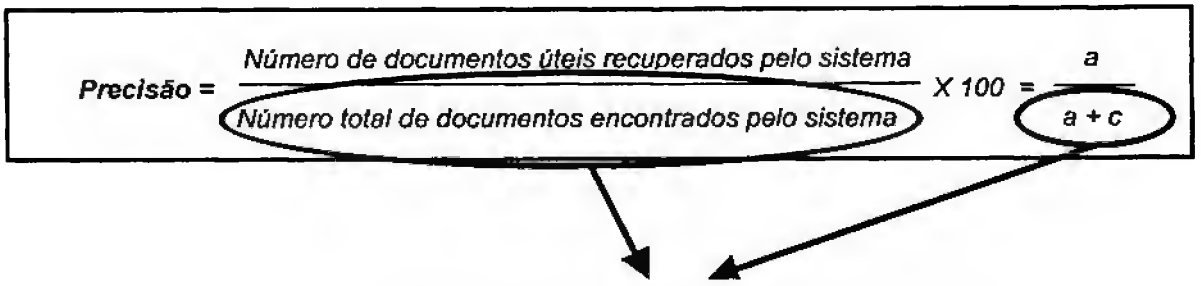
Diante da análise dos dados, chega-se a conclusão de que o primeiro pressuposto está comprovado. Assim sendo, independentemente da utilização da ferramenta de mineração de textos em relação à lista de palavras-chave utilizadas na indexação manual (Base de dados do Infohab), não há ganho significativo no índice de precisão no processo de busca e recuperação da informação.

7.2- Recuperação de itens bibliográficos

A quantidade de itens bibliográficos recuperados na Base de dados do Infohab, foi realizado a partir de 22 testes aplicados com 22 especialistas da CAIXA. Para cada teste foi calculado o número total de documentos encontrados pela Base de dados do Infohab durante o processo de busca e recuperação da informação.

O número total de documentos encontrados é dado pelo denominador da proporção utilizada para o cálculo do índice de precisão. Neste caso não há interferência dos usuários na avaliação dos resultados, pois o foco está na quantidade de itens bibliográficos recuperados.

Na fórmula utilizada para o cálculo do índice de precisão, pode-se ilustrar a posição do número total de documentos recuperados:



Cálculo do número total de documentos encontrados pela Base de dados

Onde: *a* = documentos úteis e recuperados e *c* = documentos inúteis e recuperados.

Para o cálculo do número total de documentos encontrados pela Base de dados, foram somados somente "*a*" e "*c*". Daí a importância da lista de itens bibliográficos recuperada a partir da formulação da pergunta ao sistema. Este

procedimento foi realizado com cada um dos 22 usuários que responderam à entrevista de referência.

A Tabela 13 a seguir, apresenta a quantidade de itens bibliográficos recuperados nos 22 testes realizados na Base de dados do Infohab.

Tabela 13 - Quantidade de itens bibliográficos recuperados nos 22 testes realizados na Base de dados do Infohab

Testes aplicados à Base de dados do Infohab	Quantidade de itens bibliográficos recuperados
Teste 1 (T1)	1 item
Teste 2 (T2)	6 itens
Teste 3 (T3)	2 itens
Teste 4 (T4)	1 item
Teste 5 (T5)	4 itens
Teste 6 (T6)	20 itens
Teste 7 (T7)	4 itens
Teste 8 (T8)	2 itens
Teste 9 (T9)	6 itens
Teste 10 (T10)	2 itens
Teste 11 (T11)	5 itens
Teste 12 (T12)	13 itens
Teste 13 (T13)	Nenhum item
Teste 14 (T14)	8 itens
Teste 15 (T15)	3 itens
Teste 16 (T16)	3 itens
Teste 17 (T17)	2 itens
Teste 18 (T18)	8 itens
Teste 19 (T19)	7 itens
Teste 20 (T20)	5 itens
Teste 21 (T21)	7 itens
Teste 22 (T21)	Nenhum item
Total	**109 itens bibliográficos**

A análise do Gráfico 12 relativo à quantidade de itens bibliográficos recuperados por teste na Base de dados do Infohab, mostra uma grande

dispersão quanto a quantidade de itens bibliográficos recuperados, revelando alguma regularidade apenas na faixa compreendida entre 2 e 6 itens recuperados situados abaixo do nível intermediário do Gráfico, já o máximo de itens recuperados foi de 20 itens (teste 6).

Em dois testes (T13 e T22) nenhum item bibliográfico foi recuperado. Este resultado fez com que em dois dos cálculos de precisão realizados com a Base de dados do Infohab, o índice apurado fosse nulo. Dois testes (T1 e T4), apresentaram apenas um item bibliográfico recuperado o que corresponde a 9,09%.

Depois da quantidade máxima de itens bibliográficos recuperados no teste 6 (20 itens), o de número 12 recuperou 13 itens. Curiosamente a partir daí, foram encontrados em praticamente todos os resultados, quantidades iguais de itens recuperados a cada dois testes. Assim, em dois deles (T14 e T18) foram recuperados oito itens, ou seja, 9,09% do total de vinte e dois; com sete itens recuperados, dois testes (T19 e T21), ou 9,09% do total; dois testes com seis itens recuperados (T2 e T9), ou 9,09% do total; dois testes com cinco itens (T 11 e T 20), ou 9,09%; com quatro itens recuperados, dois testes (T5 e T7), ou 9,09%; em mais dois testes (T15 e T16), três itens foram recuperados, 9,09% do total, em quatro testes (T3, T8, T10 e T17) mais dois itens foram recuperados, 18,18% do total.

Os resultados elencados dão uma conformidade irregular aos retângulos do Gráfico 12, sendo que na faixa situada entre 0 e 6 itens bibliográficos recuperados, encontra-se a maior concentração no caso da Base de dados do Infohab. São no total 16 testes que representam 72,72% de todos os vinte e dois testes realizados. Este panorama rebaixa a média de itens bibliográficos recuperados com a Base de dados do Infohab.

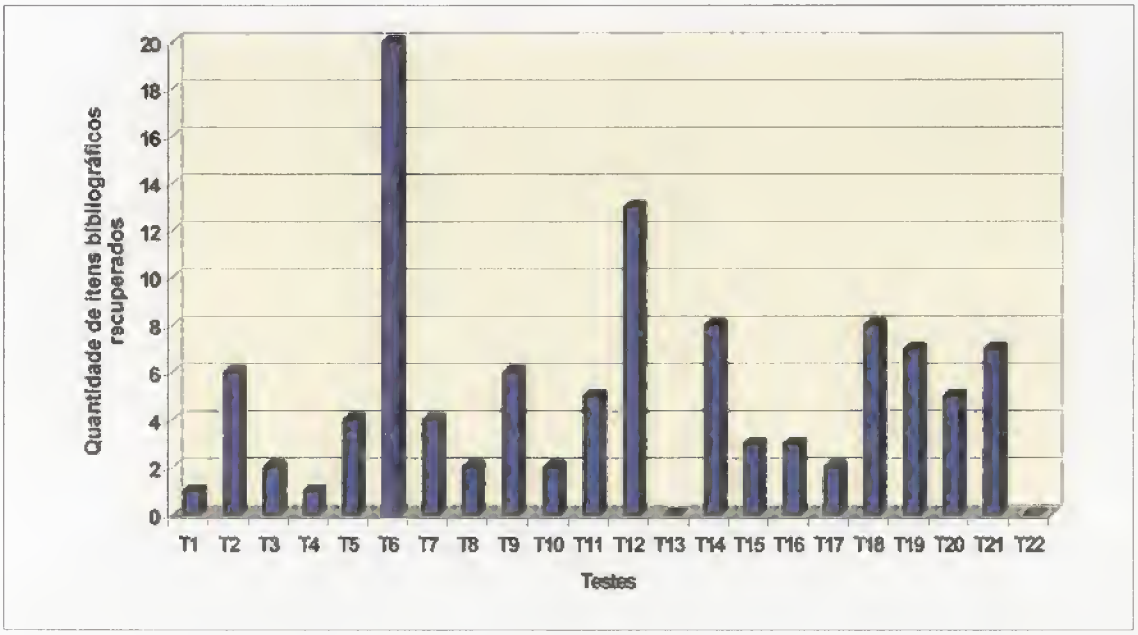


Gráfico 12 – Quantidade de itens bibliográficos recuperados por teste na Base de dados do Infohab

A Tabela 14 a seguir, apresenta a quantidade de itens bibliográficos recuperados nos 22 testes realizados no Protótipo com mineração de textos.

Tabela 14 - Quantidade de itens bibliográficos recuperados nos 22 testes realizados no Protótipo com mineração de textos

Testes aplicados à Base de dados do Infohab	Quantidade de itens bibliográficos recuperados
Teste 1 (T1)	22 itens
Teste 2 (T2)	2 itens
Teste 3 (T3)	30 itens
Teste 4 (T4)	13 itens
Teste 5 (T5)	6 itens
Teste 6 (T6)	18 itens
Teste 7 (T7)	32 itens
Teste 8 (T8)	10 itens
Teste 9 (T9)	12 itens
Teste 10 (T10)	19 itens
Teste 11 (T11)	28 itens
Teste 12 (T12)	27 itens
Teste 13 (T13)	6 itens
Teste 14 (T14)	23 itens
Teste 15 (T15)	38 itens
Teste 16 (T16)	26 itens
Teste 17 (T17)	28 itens

Teste 18 (T18)	18 itens
Teste 19 (T19)	21 itens
Teste 20 (T20)	12 itens
Teste 21 (T21)	43 itens
Teste 22 (T21)	29 itens
Total	**463 itens bibliográficos**

A análise do Gráfico 13 relativo à quantidade de itens bibliográficos recuperados por teste no Protótipo com mineração de textos apresenta uma grande dispersão. Todavia, fica claro que a quantidade de itens bibliográficos recuperados é muito superior àquela recuperada com a Base de dados do Infohab (ver Tabelas 13 e 14).

A maior quantidade de itens bibliográficos recuperados foi de quarenta e três no teste 21, e a menor quantidade foi de dois itens, encontrada no teste de número 2. Esta diferença acentuada entre a maior e a menor quantidade de itens bibliográficos recuperados dá, em certa medida, a noção da dispersão dos dados obtidos.

Cabe salientar, também, que sempre com os testes aplicados ao protótipo, foram encontrados itens bibliográficos. Ainda assim, o fato de que sempre no protótipo com mineração de textos se recuperaram itens bibliográficos em quantidades maiores do que com a Base de dados do Infohab acaba por alterar o índice médio de precisão baixo observado nos cálculos da média do índice de precisão com o Protótipo.

Depois da quantidade máxima de itens bibliográficos recuperados no teste 21 (43 itens), e da quantidade mínima de itens observada no teste 2 (dois itens), os resultados apurados foram:

- Três testes (T3, T7 e T15) no intervalo entre 30 e 38 itens recuperados, ou 13,63% do total;

- No intervalo entre 20 e 29 itens recuperados, oito testes foram encontrados (T1, T11, T12, T14, T16, T17, T19, e T22), perfazendo 36,36% dos vinte e dois testes;
- No intervalo entre 10 e 19 itens bibliográficos recuperados, sete testes (T4, T6, T8, T9, T10, T18 e T20) foram encontrados, ou 31,81% do total; e
- No intervalo entre 3 e 9 itens, dois testes (T5 e T13) foram encontrados, perfazendo 9,09% do total de testes.

Os resultados elencados dão uma conformidade irregular aos retângulos do Gráfico 13, que pode ser examinado a seguir, sendo que na faixa situada entre 20 e 29 itens bibliográficos recuperados, encontra-se a maior concentração no caso do Protótipo com mineração de textos.

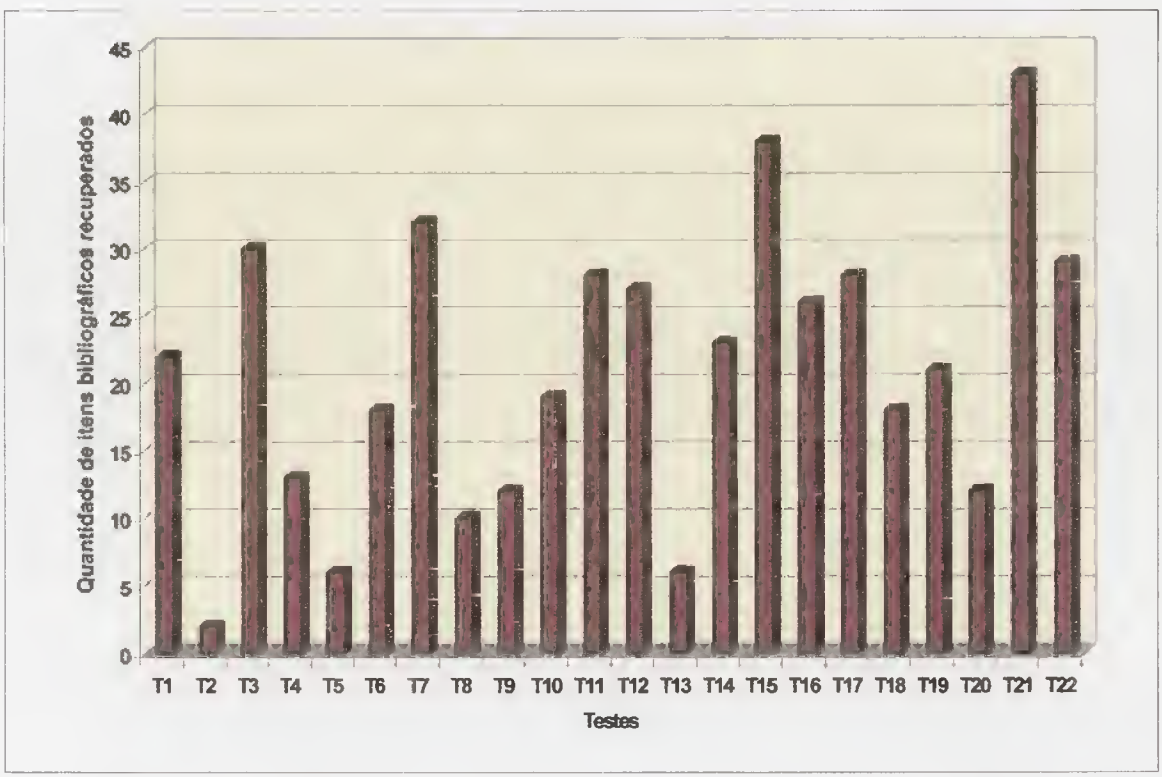


Gráfico 13 – Quantidade de itens bibliográficos recuperados por teste no Protótipo

Na Tabela 15 é apresentada a comparação das quantidades de itens bibliográficos recuperados na Base de dados do Infohab e no Protótipo com aplicação da mineração de textos, a fim de que se possa conferir em qual situação são recuperadas as maiores quantidades de itens bibliográficos.

Tabela 15 – Comparativo das quantidades de itens bibliográficos recuperados na Base de dados do Infohab e no Protótipo com mineração de textos

Testes aplicados à Base de dados do Infohab e ao Protótipo com mineração de textos	**Quantidade de itens bibliográficos recuperados na Base do Infohab**	**Quantidade de itens bibliográficos recuperados no Protótipo**
Teste 1 (T1)	1 item	22 itens
Teste 2 (T2)	6 itens	2 itens
Teste 3 (T3)	2 itens	30 itens
Teste 4 (T4)	1 item	13 itens
Teste 5 (T5)	4 itens	6 itens
Teste 6 (T6)	20 itens	18 itens
Teste 7 (T7)	4 itens	32 itens
Teste 8 (T8)	2 itens	10 itens
Teste 9 (T9)	6 itens	12 itens
Teste 10 (T10)	2 itens	19 itens
Teste 11 (T11)	5 itens	28 itens
Teste 12 (T12)	13 itens	27 itens
Teste 13 (T13)	Nenhum item	6 itens
Teste 14 (T14)	8 itens	23 itens
Teste 15 (T15)	3 itens	38 itens
Teste 16 (T16)	3 itens	26 itens
Teste 17 (T17)	2 itens	28 itens
Teste 18 (T18)	8 itens	18 itens
Teste 19 (T19)	7 itens	21 itens
Teste 20 (T20)	5 itens	12 itens
Teste 21 (T21)	7 itens	43 itens
Teste 22 (T22)	Nenhum item	29 itens
Total	**109 itens bibliográficos**	**463 itens bibliográficos**

A análise do Gráfico 14 relativo à comparação entre as quantidades de itens bibliográficos recuperados os índices de precisão na Base de dados do Infohab e no Protótipo com mineração de textos, permite inferir que a quantidade de itens bibliográficos recuperados no Protótipo é bem maior do que na Base de

dados. Com os mesmos descritores submetidos à Base de dados do Infohab e ao Protótipo com mineração de textos, a quantidade de itens recuperados na Base é de 109, enquanto que no Protótipo este número mais do quadruplica atingindo 463 itens bibliográficos recuperados.

Nos quarenta e quatro testes realizados, ou seja, vinte e dois testes na Base de dados e vinte e dois idênticos testes no protótipo, a maior quantidade recuperada na Base de dados foi de 20 itens bibliográficos, enquanto que a maior quantidade recuperada no Protótipo foi de 43 itens bibliográficos. A menor quantidade recuperada na Base de dados foi de nenhum item recuperado em dois testes (T 13 e T22), enquanto que no Protótipo foi de 2 itens em uma ocasião (T2). Cabe ainda observar que na quantidade de itens bibliográficos recuperados na Base de dados, apenas em dois testes (T6 e T12), foram encontrados números com dois dígitos, 20 e 13 respectivamente. Já na quantidade de itens bibliográficos recuperados no Protótipo, apenas em três testes (T2, T5 e T13), foram encontrados números com apenas um dígito.

Quando os resultados da quantidade de itens bibliográficos recuperados são confrontados com os resultados do cálculo do índice médio percentual de precisão realizado na Base de dados do Infohab e no protótipo, pode-se inferir que a variação dos índices de precisão é inversamente proporcional à quantidade de itens bibliográficos recuperados, ou seja, quanto mais itens bibliográficos forem recuperados (grande revocação), maior será a imprecisão da resposta; e quanto menos itens bibliográficos forem recuperados (maior especificidade), maior será a precisão. Esta constatação encontra assentimento em Foskett (1996): caso seja necessário aumentar a revocação, certamente uma parte da especificidade estará prejudicada, fato que implicará na impossibilidade de aumento da precisão.

O Gráfico 14 a seguir, ilustra as inferências feitas na comparação da quantidade de itens bibliográficos recuperados.

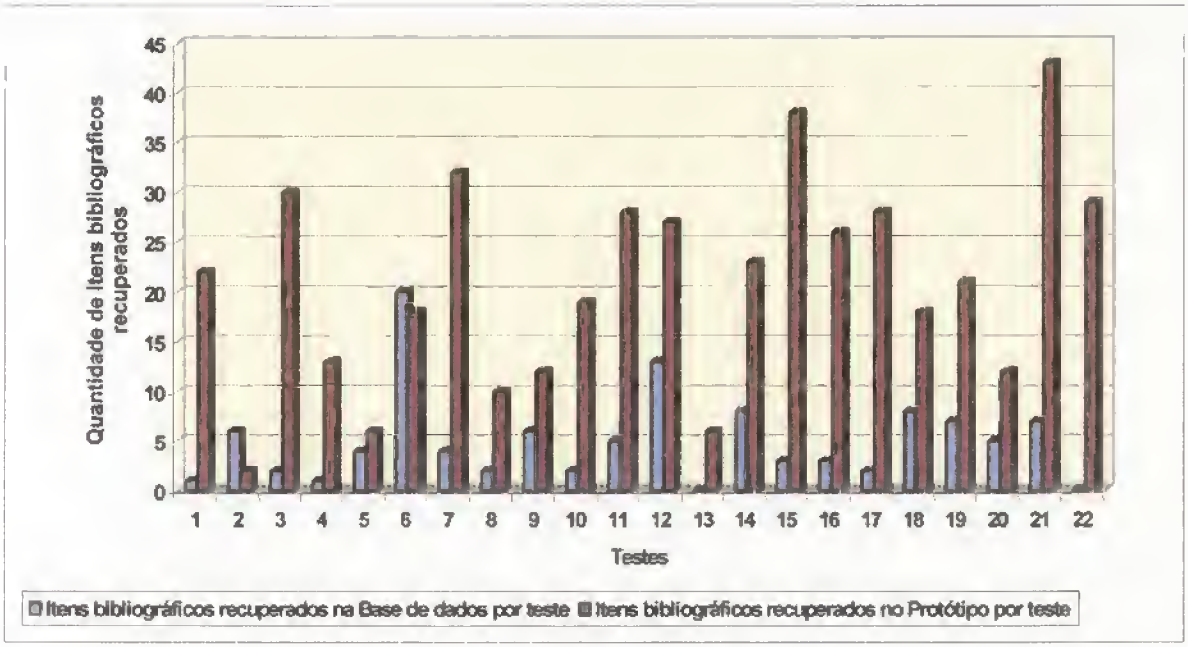


Gráfico 14 – Comparação entre a quantidade de itens bibliográficos recuperados na Base de dados do Infohab e no Protótipo

7.2.1- Comprovação do 2º pressuposto

2º Pressuposto: O uso da ferramenta de mineração de textos na busca e recuperação da informação trará sempre como resposta uma maior quantidade de itens bibliográficos do que a lista de palavras-chave utilizadas na indexação manual.

Como forma de confirmar as inferências até aqui feitas, foi calculado a partir das quantidades de itens bibliográficos recuperados da Base de dados do Infohab e do Protótipo com uso da ferramenta de mineração de textos, uma média de todos os 22 testes realizados com um e com outro.

O resultado a que se chegou, confirma as análises quanto ao fato de que a quantidade de itens bibliográficos recuperados no Protótipo é maior do que com a Base de dados. A Tabela abaixo dá uma noção clara desta diferença em termos percentuais:

Tabela 16 - Resultado comparado das médias da quantidade de itens bibliográficos recuperados na Base de dados e no Protótipo

Quantidade de itens bibliográficos recuperados na Base do Infohab	Quantidade de itens bibliográficos recuperados no Protótipo
Teste 1: 1 item	**Teste 1:** 22 itens
Teste 2: 6 itens	**Teste 2:** 2 itens
Teste 3: 2 itens	**Teste 3:** 30 itens
Teste 4: 1 item	**Teste 4:** 13 itens
Teste 5: 4 itens	**Teste 5:** 6 itens
Teste 6: 20 itens	**Teste 6:** 18 itens
Teste 7: 4 itens	**Teste 7:** 32 itens
Teste 8: 2 itens	**Teste 8:** 10 itens
Teste 9: 6 itens	**Teste 9:** 12 itens
Teste 10: 2 itens	**Teste 10:** 19 itens
Teste 11: 5 itens	**Teste 11:** 28 itens
Teste 12: 13 itens	**Teste 12:** 27 itens
Teste 13: Nenhum item	**Teste 13:** 6 itens
Teste 14: 8 itens	**Teste 14:** 23 itens
Teste 15: 3 itens	**Teste 15:** 38 itens
Teste 16: 3 itens	**Teste 16:** 26 itens
Teste 17: 2 itens	**Teste 17:** 28 itens
Teste 18: 8 itens	**Teste 18:** 18 itens
Teste 19: 7 itens	**Teste 19:** 21 itens
Teste 20: 5 itens	**Teste 20:** 12 itens
Teste 21: 7 itens	**Teste 21:** 43 itens
Teste 22: Nenhum item	**Teste 22:** 29 itens
Total: 109 itens bibliográficos Média: 4,95	**Total: 463 itens bibliográficos Média: 21,04**

A diferença entre as médias da quantidade de itens bibliográficos recuperados na Base de dados e no Protótipo é de 16,09 em favor do Protótipo. No gráfico 15 esta diferença fica bem marcada e confirma a comparação realizada teste a teste sobre a quantidade de itens recuperados.

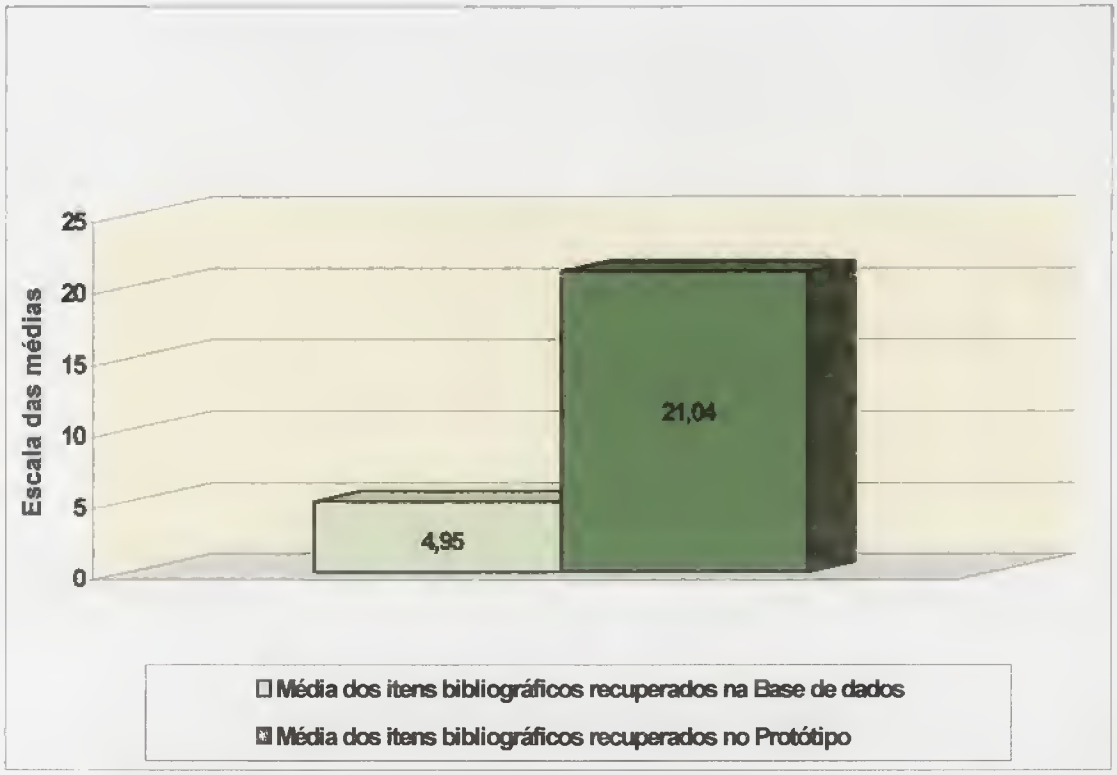


Gráfico 15 – Comparação entre as médias dos itens bibliográficos recuperados na Base de dados e no Protótipo

Finalmente, duas conclusões derivadas das análises empreendidas devem ser mais uma vez explicitadas na comprovação do segundo pressuposto da pesquisa:

I. A quantidade de itens bibliográficos recuperados no Protótipo com uso da ferramenta de mineração de textos é maior do que com a Base de dados do Infohab; e

II. Os resultados da quantidade de itens bibliográficos recuperados, quando confrontados com os resultados do cálculo do índice médio percentual de precisão realizado na Base de dados do Infohab e no protótipo, são inversamente proporcionais à quantidade de itens bibliográficos recuperados, ou seja, quanto mais itens bibliográficos forem recuperados, maior será a imprecisão da resposta.

A interpretação deste conjunto de informações comprova o segundo pressuposto. Assim, o uso da ferramenta de mineração de textos na busca e recuperação da informação trará sempre como resposta uma maior quantidade de itens bibliográficos do que a lista de palavras-chave utilizadas na indexação manual.

7.3- Recuperação de itens bibliográficos nulos

A recuperação de itens bibliográficos nulos, ou seja, nenhum item bibliográfico recuperado em um teste, acaba por anular também o índice de precisão. Quando não se recupera nenhum item bibliográfico de uma base de dados, nada pode ser submetido ao julgamento do usuário.

Como procedimento metodológico na verificação desta questão foi utilizada a somatória da quantidade de itens bibliográficos recuperados na Base de dados do Infohab e no Protótipo com uso da ferramenta de mineração de textos.

Para cada um dos 22 testes aplicados à Base e ao Protótipo, perfazendo um total de 44 testes, os mesmos descritores, com a mesma estratégia de busca, foram rigorosamente usados para um e para outro. Com isto, ficou garantida a factibilidade da pesquisa.

Em cada teste somente a somatória de todos os itens recuperados que, na fórmula do cálculo do índice de precisão está localizada no denominador da proporção, foi usada para a comprovação do segundo e terceiro pressupostos.

Esta medida é um dos elementos mais importantes para o cálculo do índice de precisão, já que a partir da lista gerada, é que os usuários realizaram o seu julgamento, anotando "útil ou inútil" para cada item bibliográfico recuperado.

A Tabela 17, a seguir, apresenta os testes realizados na Base de dados do Infohab com nenhum item bibliográfico recuperado dentre os 22 testes com os respectivos descritores e estratégias de busca.

Tabela 17 – Testes realizados na Base de dados do Infohab com nenhum item bibliográfico recuperado

Testes aplicados à Base de dados do Infohab	Descritores usados nos testes	Estratégias de busca utilizadas nos testes	Quantidade de itens bibliográficos recuperados
Teste 13 (T13)	a) Vistoria de obra; e b) Habite-se	a)Vistoria com obra; e b) Habite-se	Nenhum item
Teste 22 (T21)	a) Plano diretor; b) Risco urbano; e c) Regularização fundiária	a) Diretor com plano; b) Risco com urbano; e c) Fundiária com regularização	Nenhum item

A análise do Gráfico 16 relativo à quantidade de itens bibliográficos nulos na Base de dados do Infohab corresponde a apenas 1,83% total de itens bibliográficos recuperados.

Entretanto, ao somar estes dois resultados nulos às baixas quantidades de itens recuperados pela base de dados, entende-se a razão pela qual há uma grande disparidade entre os itens bibliográficos recuperados pela Base de dados e pelo Protótipo.

A visualização do Gráfico 16, permite também inferir que a resposta na Base de dados, considerando a possibilidade de ampliação da quantidade de testes a serem aplicados, possivelmente poderá ampliar, ainda mais, a quantidade de testes com nenhum item bibliográfico recuperado.

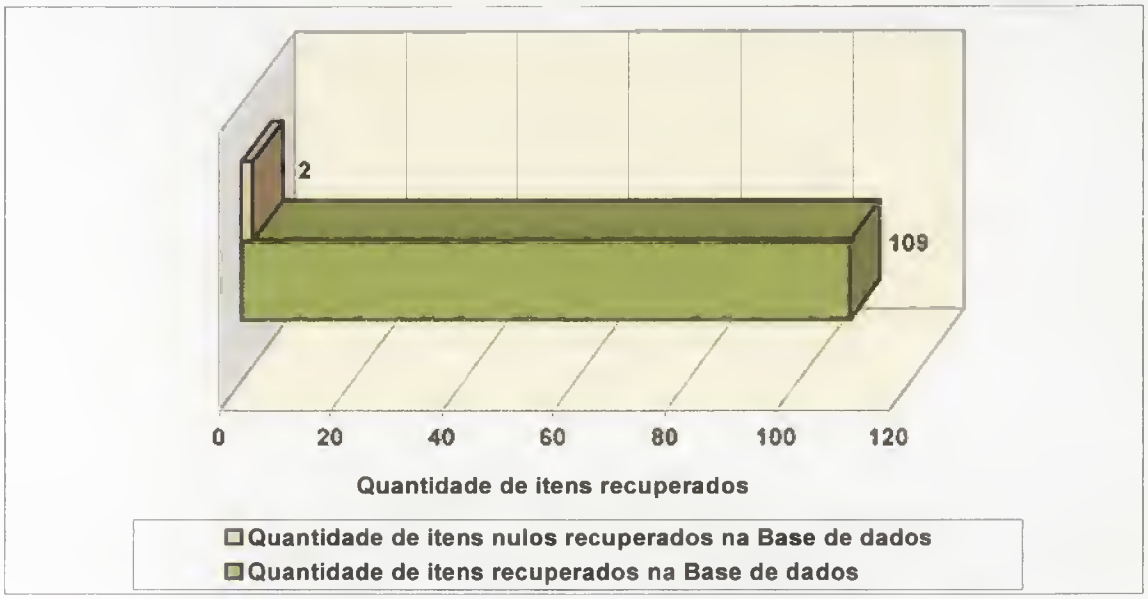


Gráfico 16 – Quantidade de itens nulos recuperados na Base de dados do Infohab

Já no Protótipo com uso da ferramenta de mineração de textos (Gráfico 17), não houve ocorrência de itens bibliográficos nulos. A menor quantidade de itens encontrados foi de 2 itens no teste de número 2 e a maior foi de quarenta e três itens bibliográficos recuperados no teste de número 21. Nos demais os resultados obtidos foram:

- Três testes (T3, T7 e T15) no intervalo entre 30 e 38 itens recuperados, ou 13,63% do total;
- No intervalo entre 20 e 29 itens recuperados, oito testes foram encontrados (T1, T11, T12, T14, T16, T17, T19, e T22), perfazendo 36,36% dos vinte e dois testes;
- No intervalo entre 10 e 19 itens bibliográficos recuperados, sete testes (T4, T6, T8, T9, T10, T18 e T20) foram encontrados, ou 31,81% do total; e

- No intervalo entre 3 e 9 itens, dois testes (T5 e T13) foram encontrados, perfazendo 9,09% do total de testes.

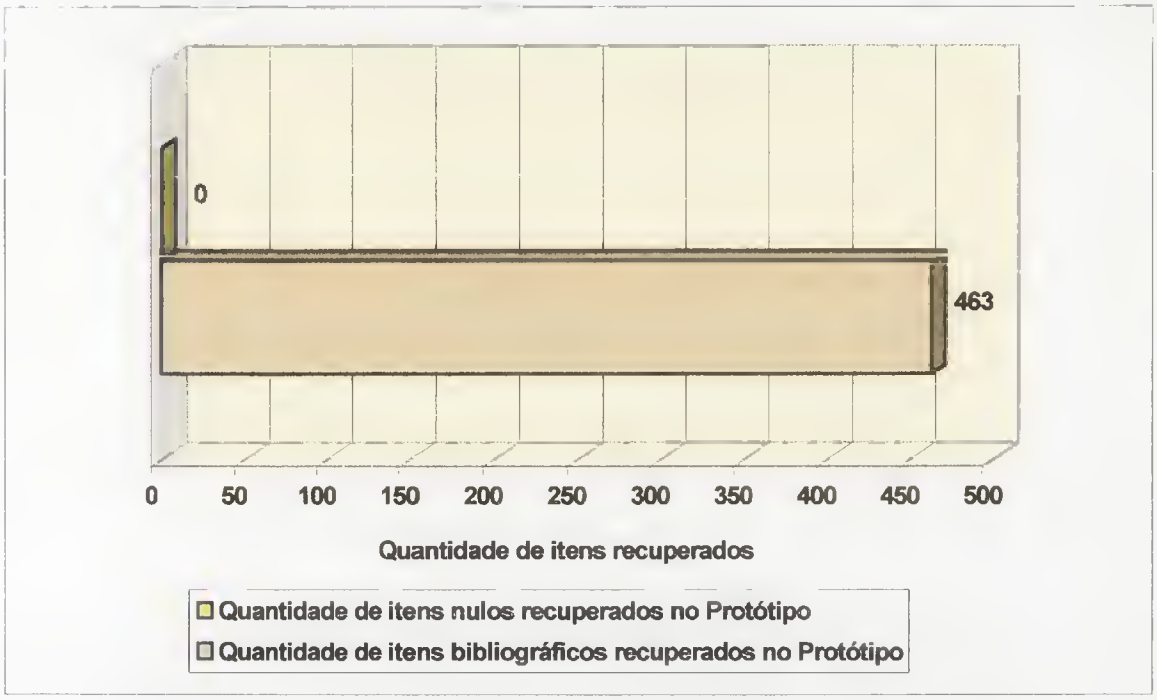


Gráfico 17 – Quantidade de itens bibliográficos nulos no Protótipo com uso da ferramenta de mineração de textos

A Tabela 18, a seguir, apresenta os testes nulos quanto à quantidade de itens bibliográficos na Base de dados do Infohab com os respectivos descritores e estratégias de busca, comparados aos mesmos testes aplicados ao protótipo com uso da ferramenta de mineração de textos.

Tabela 18 – Comparativo entre os testes realizados na Base de dados do Infohab com nenhum item bibliográfico recuperado e os mesmos testes no Protótipo

Testes aplicados à Base de dados e ao Protótipo	Descritores usados nos testes	Estratégias de busca utilizadas nos testes	Quantidade de itens bibliográficos recuperados na Base de dados	Quantidade de itens bibliográficos recuperados no Protótipo
Teste 13 (T13)	a) Vistoria de obra; e b) Habite-se	a)Vistoria com obra; e b) Habite-se	Nenhum item	6 itens
Teste 22 (T22)	a) Plano diretor; b) Risco urbano; e c) Regularização fundiária	a) Diretor com plano; b) Risco com urbano; e c) Fundiária com regularização	Nenhum item	29 itens

A Comparação entre as quantidades de itens bibliográficos nulos recuperados na Base de dados do Infohab e no Protótipo com uso da ferramenta de mineração de textos pode ser melhor cotejada no Gráfico 18, a seguir:

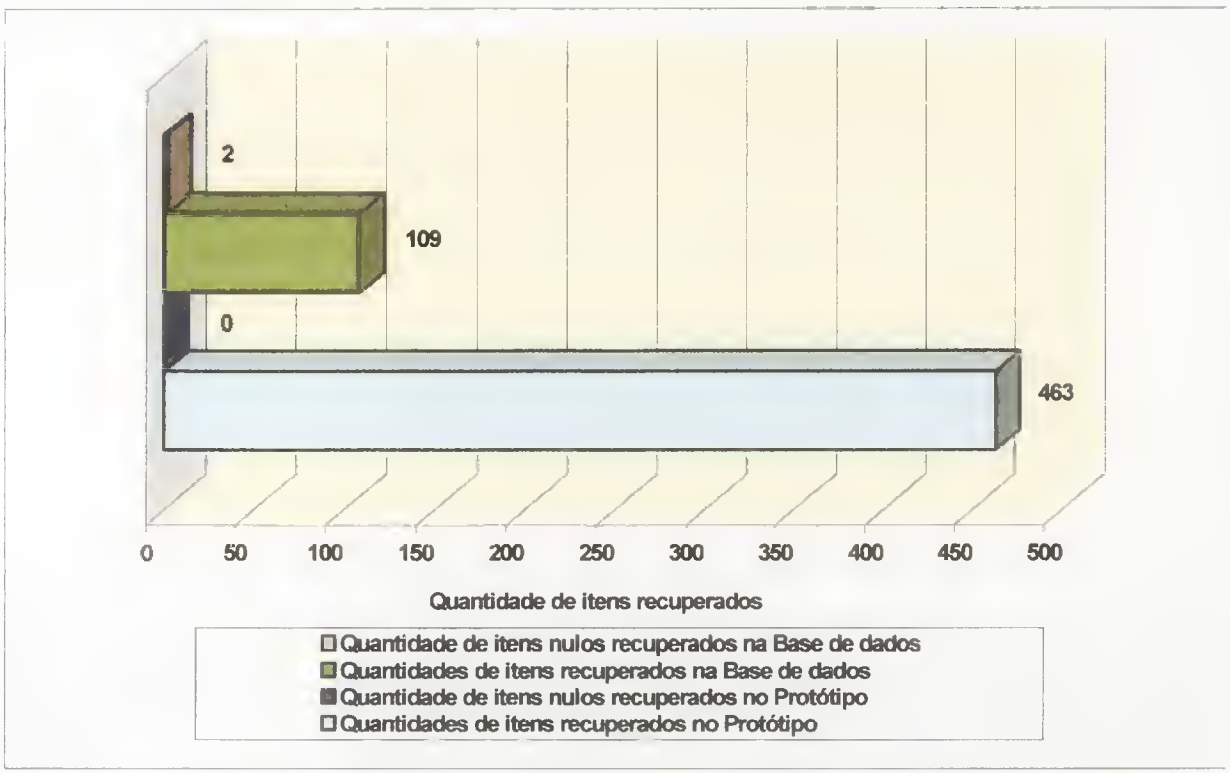


Gráfico 18 – Comparação entre as quantidades de itens bibliográficos nulos recuperados na Base de dados do Infohab e no Protótipo

7.3.1- Comprovação do 3º pressuposto

3º Pressuposto: Quando termos específicos da Base de dados do Infohab são submetidos à lista de palavras-chave utilizadas na indexação manual na busca e recuperação da informação, nem sempre são encontrados itens bibliográficos, enquanto que os mesmos termos quando aplicados ao Protótipo com aplicação da mineração de textos, irão recuperar algum item bibliográfico.

A comprovação do terceiro pressuposto demonstra que o uso da ferramenta de mineração de textos na busca e recuperação da informação trará como resposta uma maior quantidade de itens bibliográficos do que a lista de palavras-chave utilizadas na indexação manual, além de não apresentar resultados nulos.

Na Base de dados do Infohab, a tendência é contrária, ou seja, é possível inferir que a resposta na Base de dados, considerando a possibilidade de ampliação da quantidade de testes a serem submetidos, possivelmente poderá ampliar, ainda mais, a quantidade de testes com nenhum item bibliográfico recuperado.

Há uma grande potencialidade de recuperação de itens em qualquer parte do texto completo ao se utilizar a extração automática de termos, o que explica a comprovação do terceiro pressuposto, já que a ferramenta de mineração de textos captura qualquer termo em qualquer parte do texto completo armazenado no Protótipo.

A interpretação deste conjunto de informações, então, comprova o terceiro pressuposto. Assim, quando termos específicos da Base de dados do Infohab são submetidos à lista de palavras-chave utilizadas na indexação manual na busca e recuperação da informação, nem sempre são encontrados itens bibliográficos, enquanto que os mesmos termos quando são submetidos ao Protótipo com aplicação da mineração de textos, encontra sempre algum item bibliográfico.

7.4- Resultados preliminares

A comprovação dos três pressupostos analisados respondeu às três questões propostas na seção 1.2- Propósito da pesquisa (p. 6).

A primeira questão indaga se a mineração de textos aplicada ao processo de busca e recuperação da informação traz ganhos de precisão se comparada à indexação manual. Os dados obtidos comprovaram o 1º pressuposto e revelaram que não há ganho significativo no índice de precisão no processo de busca e recuperação da informação.

A segunda questão indaga se a mineração de textos pode ser empregada como ferramenta complementar no processo de indexação, visando o aumento do índice de precisão na recuperação da informação. Os dados coletados e analisados que comprovam os pressupostos 2 e 3 mostraram que o uso da ferramenta de mineração de textos na busca e recuperação da informação trará sempre como resposta uma maior quantidade de itens bibliográficos do que a lista de palavras-chave utilizadas na indexação manual; já a submissão de termos específicos da Base de dados do Infohab ao Protótipo com aplicação da mineração de textos, sempre trará itens bibliográficos. Tais confirmações abrem caminho para uma resposta positiva à segunda indagação e, por extensão, à terceira que questiona se é possível a construção de uma sistemática de uso de mineração de textos para complementar e aperfeiçoar o processo de indexação, visando o aumento do índice de precisão na recuperação da informação.

Considerando que há factibilidade para responder positivamente as duas últimas questões, as premissas do problema (seção 1.3- Premissas básicas) podem ser consideradas como verdadeiras.

O uso da 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, desde que o julgamento do indexador seja considerado indispensável na montagem de uma sistemática de uso dos termos gerados a partir da mineração de textos. A habilidade do indexador de contextualizar, relacionar palavras, usar a abstração, bem como decidir quais termos serão usados para identificar o conteúdo dos documentos, é fator preponderante para a indexação de documentos e a sua posterior recuperação em uma base de dados.

Enfim, considerando os argumentos que comprovam o segundo e terceiro pressupostos, conclui-se que a mineração de textos (utilizando-se do *Software* BR/*Search* na criação do Protótipo) trará como resposta a uma consulta, uma quantidade de itens bibliográficos sempre maior do que na resposta obtida com a lista de palavras-chave utilizadas na indexação manual.

Esta possibilidade, com o reforço da verificação do 1º pressuposto, abre caminho para o emprego da mineração de textos como instrumento de enriquecimento da lista de palavras-chave e/ou construção de um vocabulário controlado utilizando, para tanto, a lista de palavras mais freqüentes em cada documento (descoberta por listas de conceitos-chave) recuperada em pesquisas a serem realizadas no protótipo, além da lista de palavras mais freqüentes do resultado total da pesquisa (descoberta por descrição de classes de textos), também a ser realizada no protótipo. Faz-se necessário ressaltar, neste contexto, que a grande potencialidade da ferramenta é a captura de qualquer termo em qualquer parte do texto completo armazenado no Protótipo. O que se configura como um instrumento útil no aprimoramento contínuo do processo de indexação, já que a mineração de textos pode extrair automaticamente termos relacionados com a pesquisa (resultado da entrevista de referência). A partir daí, o julgamento dos indexadores, que deverão selecionar os termos a serem usados na representação do conteúdo dos documentos, poderá enriquecer e/ou apoiar a construção de um tesauro, por exemplo.

Os resultados obtidos na comprovação dos pressupostos 1, 2 e 3 da pesquisa são determinantes na formulação do argumento de que a mineração de textos pode ser utilizada também como instrumento de enriquecimento da lista de palavras-chave e/ou construção de um vocabulário controlado, utilizando a lista de palavras mais freqüentes em cada documento.

Duas são as possibilidades concretas da mineração de textos por meio do *Software* BR/*Search*:

- A geração e utilização da lista de palavras mais freqüentes em cada item bibliográfico (texto completo) recuperada na pesquisa realizada no protótipo; e

- A geração e utilização da lista de palavras mais freqüentes no resultado total (textos completos agrupados) da pesquisa realizada no protótipo com uso de mineração de textos.

A seguir serão apresentados exemplos de pesquisas e geração de listas de palavras mais freqüentes em cada documento (descoberta por listas de conceitos-chave) recuperadas na pesquisa realizada no protótipo e listas de palavras mais freqüentes do resultado total da pesquisa (descoberta por descrição de classes de textos) realizada no protótipo.

1º Exemplo: Termo pesquisado: revitalização COM[17] urbana com dois itens bibliográficos recuperados.

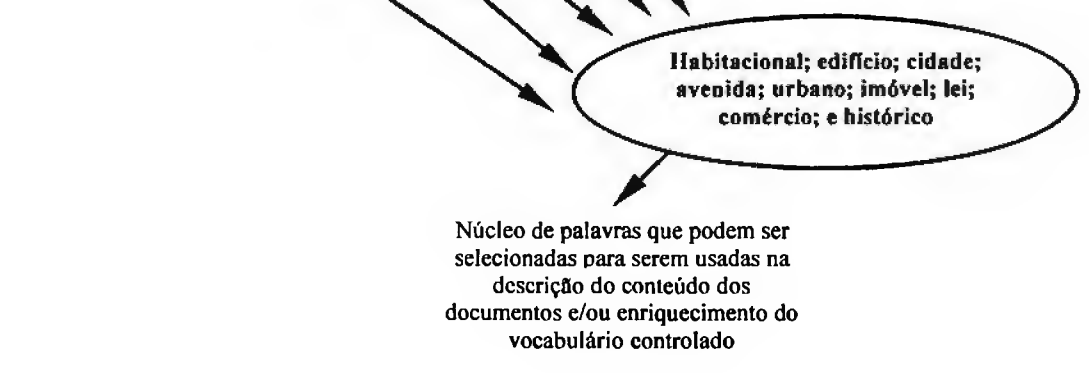
A tabela 19 apresenta a lista das 20 (vinte) palavras mais freqüentes do resultado total da pesquisa (descoberta por descrição de classes de textos, ver página 67):

[17] **COM** – Operador booleano que combina dois ou mais termos em uma expressão de busca. Na recuperação da informação no protótipo o operador booleano "com" foi empregado com a mesma função do operador "E", que executa a operação de intersecção entre conjuntos.

Tabela 19 - Lista de palavras mais freqüentes do resultado total da pesquisa realizada no Protótipo com o termo de busca: revitalização COM urbana

Palavras	Ocorrências	Ocorrências na base	Porcentagem	Documentos	Documentos na base	Porcentagem
Recife	74	308	24,02	2	10	20
área	69	1082	6,37	2	46	4,34
uso	66	624	10,57	2	49	4,08
centro	61	259	23,55	2	37	5,40
áreas	47	813	5,78	2	45	4,44
habitacional	**44**	**707**	**6,22**	**2**	**27**	**7,4**
edifícios	**42**	**121**	**34,71**	**1**	**12**	**8,33**
programa	41	777	5,27	2	44	4,54
cidade	**38**	**817**	**4,65**	**2**	**44**	**4,54**
avenida	**37**	**50**	**74**	**1**	**9**	**11,11**
cidades	**37**	**336**	**11,01**	**2**	**39**	**5,12**
urbana	**36**	**640**	**5,62**	**2**	**41**	**4,87**
edifício	**33**	**157**	**21,01**	**2**	**8**	**12,5**
imóveis	**32**	**522**	**6,13**	**2**	**24**	**8,33**
onde	32	542	5,87	2	51	3,92
Pernambuco	32	54	59,25	1	3	33,33
lei	**31**	**491**	**6,31**	**2**	**33**	**6,06**
tem	31	865	3,58	2	53	3,77
Caixa	30	669	4,48	2	36	5,55
Comércio	**30**	**86**	**34,88**	**2**	**19**	**10,52**
foi	30	995	3,01	2	53	3,77
Históricos	**30**	**78**	**38,46**	**2**	**12**	**16,66**
Urbano	**30**	**664**	**4,51**	**2**	**44**	**4,54**
Guararapes	29	34	85,29	1	2	50
Estudo	26	445	5,84	2	45	4,44

(Fonte: protótipo com aplicação de mineração de textos - software BR/*Search*)



Núcleo de palavras que podem ser selecionadas para serem usadas na descrição do conteúdo dos documentos e/ou enriquecimento do vocabulário controlado

A tabela 20 apresenta a lista das 20 (vinte) palavras mais freqüentes do primeiro documento recuperado (descoberta por listas de conceitos-chave, ver página 67):

- **Referência do documento:**

CAVALCANTE, C. Q. B. *O uso misto na reocupação dos edifícios subutilizados nos centros urbanos*. Recife, 2004.

Tabela 20 - Lista de palavras mais freqüentes do primeiro documento recuperado da pesquisa realizada no Protótipo

Palavras	Ocorrências	Ocorrências na base	Porcentagem	Documentos	Documentos na base	Porcentagem
Recife	71	308	23,05	1	10	10
uso	64	624	10,26	1	49	2,04
área	63	1082	5,82	1	46	2,17
centro	57	259	22	1	37	2,70
edifícios	**42**	**121**	**34,71**	**1**	**12**	**8,33**
habitacional	**40**	**707**	**5,65**	**1**	**27**	**3,70**
áreas	38	813	4,67	1	45	2,22
avenida	**37**	**50**	**74**	**1**	**9**	**11,11**
cidade	**37**	**817**	**4,52**	**1**	**44**	**2,27**
edifício	**33**	**157**	**21,01**	**1**	**8**	**12,5**
Pernambuco	32	54	59,25	1	3	33,33
lei	**30**	**491**	**6,10**	**1**	**33**	**3,03**
cidades	**29**	**336**	**8,63**	**1**	**39**	**2,56**
Guararapes	29	34	85,29	1	2	50
onde	29	545	5,32	1	51	1,96
comércio	**28**	**86**	**32,55**	**1**	**19**	**5,26**
tem	26	865	3	1	53	1,88
serviços	25	866	2,88	1	44	2,27
unidades	25	385	6,49	1	38	2,63
estudo	24	445	5,39	1	45	8,22
imóveis	**24**	**522**	**4,59**	**1**	**24**	**4,16**
pavimentos	24	70	34,28	1	7	14,28
térreo	24	32	75	1	2	50
urbano	**24**	**664**	**3,61**	**1**	**44**	**2,27**
foi	23	995	2,31	1	53	1,88

(Fonte: protótipo com aplicação de mineração de textos - software BR/*Search*)

Núcleo de palavras que podem ser selecionadas para serem usadas na descrição do conteúdo dos documentos e/ou enriquecimento do vocabulário controlado

A tabela 21 apresenta a lista das 20 (vinte) palavras mais freqüentes do segundo documento recuperado (descoberta por listas de conceitos-chave, ver página 67):

- **Referência do documento:**

GALIZA, H. R. S. *A reabilitação urbana dos sítios históricos brasileiros.* Rio de Janeiro, 2002.

Tabela 21 - Lista de palavras mais freqüentes do segundo documento recuperado da pesquisa realizada no Protótipo

Palavras	Ocorrências	Ocorrências na base	Porcentagem	Documentos	Documentos na base	Porcentagem
programa	22	777	2,83	1	44	2,27
históricos	**21**	**78**	**26,92**	**1**	**12**	**8,33**
reabilitação	**19**	**31**	**61,29**	**1**	**7**	**14,28**
Caixa	17	669	2,54	1	36	2,77
revitalização	**16**	**52**	**30,76**	**1**	**7**	**14,28**
sítios	**14**	**60**	**23,33**	**1**	**7**	**14,28**
patrimônio	**13**	**123**	**10,56**	**1**	**17**	**5,88**
urbana	**13**	**640**	**2,03**	**1**	**41**	**2,43**
projetos	11	594	1,85	1	42	2,38
recursos	10	1013	0,98	1	45	2,22
áreas	9	813	1,10	1	45	2,22
cidades	**8**	**336**	**2,38**	**1**	**39**	**2,56**
desenvolvimento	8	1020	0,78	1	52	1,92
financeiros	8	144	5,55	1	31	3,22
históricos	**8**	**117**	**6,83**	**1**	**27**	**3,70**
imóveis	**8**	**522**	**1,53**	**1**	**24**	**4,16**
também	8	833	0,96	1	48	2,08
técnicos	8	224	3,57	1	31	3,22
centros	7	105	6,66	1	22	4,54
cultural	7	111	6,30	1	29	3,44
foi	7	995	0,70	1	53	1,88
IPHAN[18]	**7**	**16**	**43,75**	**1**	**3**	**33,33**
outras	7	466	1,50	1	48	2,08
recuperação	7	142	4,92	1	30	3,33
área	6	1082	0,55	1	46	2,17

(Fonte: protótipo com aplicação de mineração de textos - software BR/*Search*)

[18] **IPHAN** – Instituto do Patrimônio Histórico e Artístico Nacional.

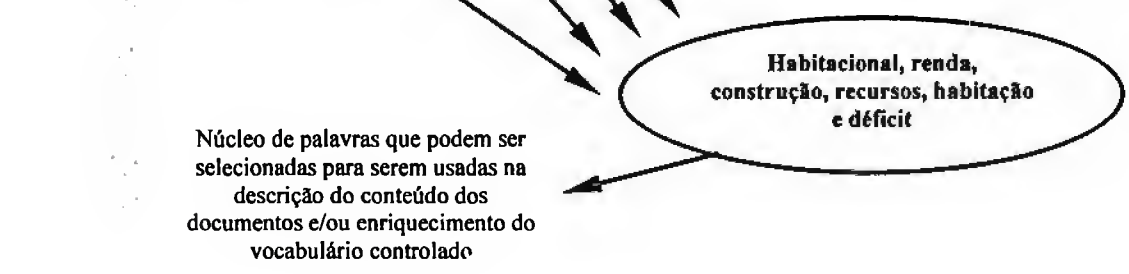
2º Exemplo: Termo pesquisado: arrendamento COM urbano três itens bibliográficos recuperados.

A tabela a seguir apresenta a lista das 20 (vinte) palavras mais freqüentes do resultado total da pesquisa (descoberta por descrição de classes de textos, ver página 67):

Tabela 22 - Lista de palavras mais freqüentes do resultado total da pesquisa Realizada no Protótipo com o termo de busca: arrendamento COM urbano

Palavras	Ocorrências	Ocorrências na base	Porcentagem	Documentos	Documentos na base	Porcentagem
to	307	948	32,38	2	31	6,45
link	305	803	37,98	2	15	13,33
page	305	803	37,98	2	15	13,33
habitacional	**274**	**707**	**38,75**	**3**	**27**	**11,11**
renda	**202**	**974**	**20,73**	**3**	**39**	**7,69**
construção	**181**	**547**	**33,08**	**3**	**46**	**6,52**
programa	179	777	23,03	3	44	6,81
recursos	**177**	**1013**	**17,47**	**3**	**45**	**6,66**
habitação	**157**	**702**	**22,36**	**3**	**33**	**9,09**
"1"	157	6953	2,25	3	55	5,45
par	156	169	92,30	3	13	23,07
população	**148**	**1054**	**14,04**	**3**	**50**	**6**
pela	144	1339	10,75	3	52	5,76
através	143	880	16,25	3	48	6,25
déficit	**137**	**274**	**50**	**3**	**18**	**16,66**
R	136	1501	9,06	3	41	7,31
"3"	134	4612	2,90	3	55	5,45
área	130	1082	12,01	3	46	6,52
"2"	124	5693	2,17	3	55	5,45
alianças	123	126	97,61	1	3	33,33
foi	123	995	12,36	3	53	5,66
"5"	120	4048	2,96	3	53	5,66
habitacionais	**119**	**295**	**40,33**	**3**	**28**	**10,71**
pesquisa	118	535	22,05	3	43	6,97
tem	117	865	13,52	3	53	5,66

(Fonte: protótipo com aplicação de mineração de textos - software BR/*Search*)



A tabela 23 apresenta a lista das 20 (vinte) palavras mais freqüentes do primeiro documento recuperado (descoberta por listas de conceitos-chave, ver página 67):

- **Referência do documento:**

CAVALCANTE, C. Q. B. *O uso misto na reocupação dos edifícios subutilizados nos centros urbanos*. Recife, 2004.

Tabela 23 - Lista de palavras mais freqüentes do primeiro documento Recuperado da pesquisa realizada no Protótipo

Palavras	Ocorrências	Ocorrências na base	Porcentagem	Documentos	Documentos na base	Porcentagem
Recife	71	308	23,05	1	10	10
uso	64	624	10,26	1	49	2,04
área	63	1082	5,82	1	46	2,17
centro	57	259	22	1	37	2,70
edifícios	42	121	34,71	1	12	8,33
habitacional	**40**	**707**	**5,65**	**1**	**27**	**3,70**
áreas	38	813	4,67	1	45	2,22
avenida	37	50	74	1	9	11,11
cidade	**37**	**817**	**4,52**	**1**	**44**	**2,27**
edifício	33	157	21,01	1	8	12,5
Pernambuco	32	54	59,25	1	3	33,33
lei	**30**	**491**	**6,10**	**1**	**33**	**3,03**
cidades	**29**	**336**	**8,63**	**1**	**39**	**2,56**
Guararapes	29	34	85,29	1	2	50
onde	29	545	5,32	1	51	1,96
comércio	**28**	**86**	**32,55**	**1**	**19**	**5,26**
tem	26	865	3	1	53	1,88
serviços	25	866	2,88	1	44	2,27
unidades	25	385	6,49	1	38	2,63
estudo	24	445	5,39	1	45	8,22
imóveis	**24**	**522**	**4,59**	**1**	**24**	**4,16**
pavimentos	**24**	**70**	**34,28**	**1**	**7**	**14,28**
térreo	24	32	75	1	2	50
urbano	**24**	**664**	**3,61**	**1**	**44**	**2,27**
foi	23	995	2,31	1	53	1,88

(Fonte: protótipo com aplicação de mineração de textos - software BR/*Search*)

Habitacional, cidade, lei, comércio, imóvel, pavimentos e urbano

Núcleo de palavras que podem ser selecionadas para serem usadas na descrição do conteúdo dos documentos e/ou enriquecimento do vocabulário controlado

A tabela 24 apresenta a lista das 20 (vinte) palavras mais freqüentes do segundo documento recuperado (descoberta por listas de conceitos-chave, ver página 67):

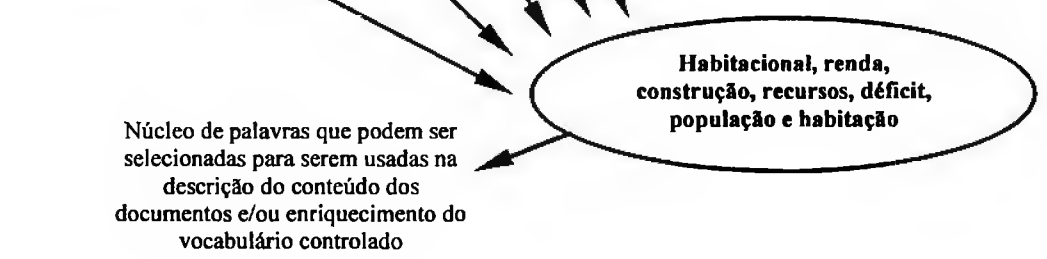
- **Referência do documento:**

MARQUES, J. A. V. *A utilização de alianças estratégicas no combate do déficit habitacional: um estudo de caso do conjunto habitacional "Cidade de Deus" Sete Lagoas - MG*. Varginha, 2002.

Tabela 24 - Lista de palavras mais freqüentes do segundo documento recuperado da pesquisa realizada no Protótipo

Palavras	Ocorrências	Ocorrências na base	Porcentagem	Documentos	Documentos na base	Porcentagem
habitacional	**170**	**707**	**24,04**	**1**	**27**	**3,70**
renda	**145**	**974**	**14,88**	**1**	**39**	**2,56**
alianças	123	126	97,61	1	3	33,33
"1"	123	6953	1,76	1	55	1,81
construção	**113**	**547**	**20,65**	**1**	**46**	**2,17**
recursos	**109**	**1013**	**10,76**	**1**	**45**	**2,22**
pesquisa	105	535	19,62	1	43	2,32
TO	103	948	10,86	1	31	3,22
R	102	1501	6,79	1	41	2,43
déficit	**101**	**274**	**36,86**	**1**	**18**	**5,55**
link	101	803	12,57	1	15	6,66
page	101	803	12,57	1	15	6,66
pela	97	1339	7,24	1	52	1,92
aliança	95	97	97,93	1	3	33,33
total	93	95	13,63	1	44	2,27
foi	90	93	9,04	1	53	1,88
estratégica	89	103	86,40	1	9	11,11
estratégicas	89	100	89	1	6	16,66
"5"	89	4048	2,19	1	53	1,88
população	**88**	**1054**	**8,34**	**1**	**50**	**2**
3	88	4612	1,90	1	55	1,81
estudo	85	445	19,10	1	45	2,22
"2"	84	5693	1,47	1	55	1,81
desenvolvimento	82	1020	8,03	1	52	1,92
habitação	**81**	**702**	**11,5**	**1**	**33**	**3,03**

(Fonte: protótipo com aplicação de mineração de textos - software BR/*Search*)



Núcleo de palavras que podem ser selecionadas para serem usadas na descrição do conteúdo dos documentos e/ou enriquecimento do vocabulário controlado

A tabela 25 apresenta a lista das 20 (vinte) palavras mais freqüentes do terceiro documento recuperado (descoberta por listas de conceitos-chave, ver página 67):

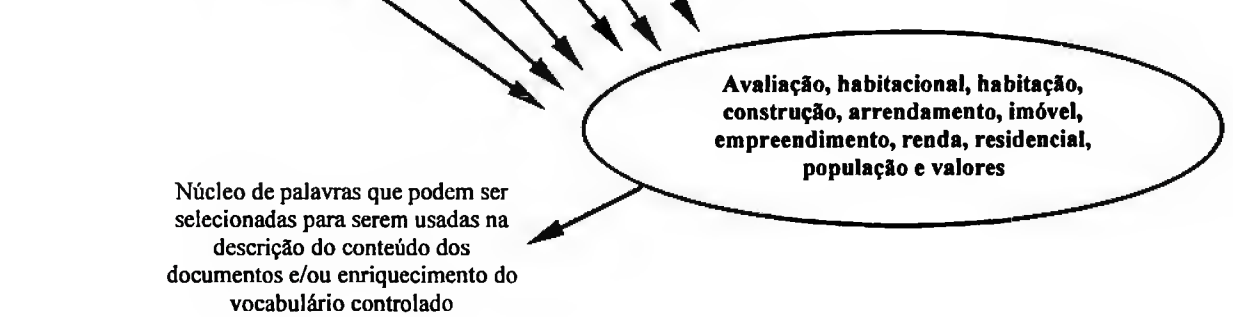
- **Referência do documento:**

LIMA, L. F. *Análise do potencial de aplicação da avaliação pós–ocupação no programa de arrendamento residencial.* Curitiba, 2002.

Tabela 25 - Lista de palavras mais freqüentes do terceiro documento recuperado da pesquisa realizada no Protótipo

Palavras	Ocorrências	Ocorrências na base	Porcentagem	Documentos	Documentos na base	Porcentagem
link	204	803	25,40	1	15	6,66
page	204	803	25,40	1	15	6,66
to	204	948	21,51	1	31	3,22
par	136	169	80,47	1	13	7,69
programa	105	777	13,51	1	44	2,27
avaliação	**93**	**511**	**18,19**	**1**	**36**	**2,77**
através	70	880	7,95	1	48	2,08
APO	66	67	98,50	1	2	50
habitacional	**64**	**707**	**9,05**	**1**	**27**	**3,70**
habitação	**58**	**702**	**8,26**	**1**	**33**	**3,03**
construção	**57**	**547**	**10,42**	**1**	**46**	**2,17**
arrendamento	**56**	**91**	**61,53**	**1**	**9**	**11,11**
problemas	56	503	11,13	1	45	2,22
imóvel	**53**	**390**	**13,58**	**1**	**20**	**5**
recursos	53	1013	5,23	1	45	2,22
empreendimento	**50**	**141**	**35,46**	**1**	**22**	**4,54**
renda	**46**	**874**	**4,72**	**1**	**39**	**2,56**
habitacionais	**44**	**295**	**14,91**	**1**	**28**	**3,57**
aplicação	43	249	17,26	1	38	2,63
áreas	43	813	5,28	1	45	2,22
residencial	**42**	**116**	**36,20**	**1**	**16**	**6,25**
análise	41	508	8,07	1	48	2,08
Caixa	40	669	5,97	1	30	2,77
população	**40**	**1054**	**3,79**	**1**	**50**	**2**
valores	**39**	**761**	**5,12**	**1**	**43**	**2,32**

(Fonte: protótipo com aplicação de mineração de textos - software BR/*Search*)



Assim, poderá ser desenhada a sistemática de uso da mineração de textos, a partir das listas de palavras mais freqüentes para apoiar o processo de indexação manual, visando o aumento do índice de precisão de resposta no processo de recuperação da informação, conforme é apresentado no fluxograma a seguir:

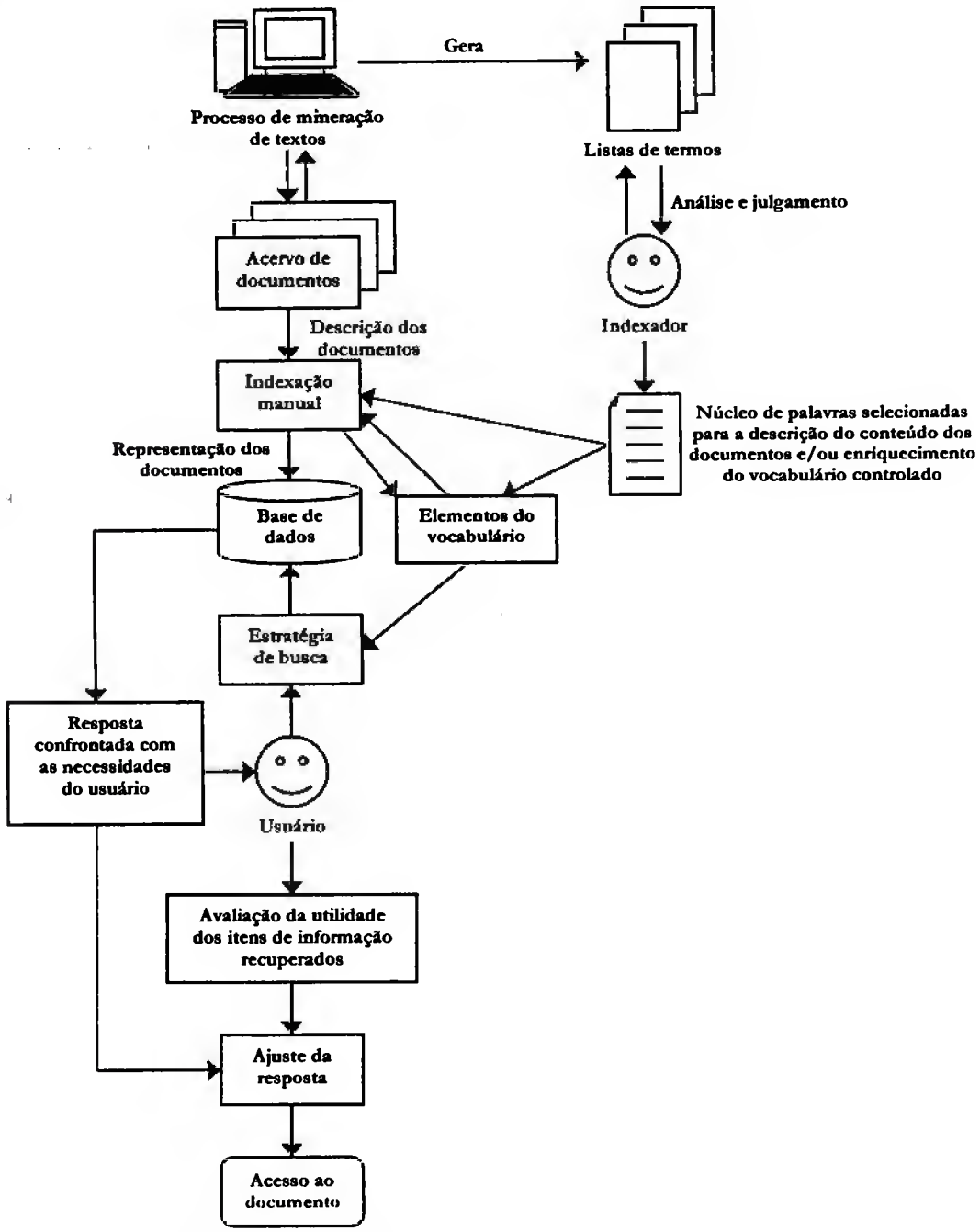


Figura 31 – Fluxograma da sistemática de uso da mineração de textos na indexação manual

8- DISCUSSÃO DAS TESES

Tese I – O valor do índice de precisão resultante do processo de busca e recuperação da informação baseado na indexação manual, não é superado pelo valor correspondente ao uso da ferramenta de mineração de textos.

A primeira tese defendida está apoiada na argumentação de que o trabalho intelectual na escolha dos termos que devem representar o conteúdo de documentos é o principal instrumento de qualquer processo de indexação manual no âmbito dos sistemas de informação. Esta situação comprova que tais sistemas, apesar de todo o avanço no processamento dos dados e o leque cada vez maior de tecnologias da informação voltadas para o seu gerenciamento, não foram capazes ainda de realizar a indexação automática em toda a sua totalidade com a mesma efetividade da indexação manual. Entretanto o uso combinado da indexação manual e da indexação automática, poderá melhorar a performance do processo de indexação, equilibrando a relação custo-benefício.

A indexação automática é, na realidade, uma combinação "automática" entre termos usados como entrada e os termos que coincidem com eles nos documentos, não incluindo aspectos de julgamento prévio dos indexadores, caros à indexação manual.

Assim, a indexação manual com todas as falhas humanas de interpretação e subjetividade que acabam sendo arbitrárias na escolha do melhor descritor, continua a ter um papel central, já consagrado na prática e na vasta literatura que trata do tema, para a descrição e representação do conteúdo dos documentos que compõem uma base de dados.

A partir da comprovação do primeiro pressuposto da pesquisa de que não há ganho significativo no índice de precisão com o uso da mineração de textos em relação à indexação manual, a primeira tese deve ser amplamente considerada ao

se adotar entusiasticamente tecnologias de mineração de textos como solução para os problemas de representação do conteúdo dos documentos, e conseqüentemente para o processo de busca e recuperação da informação. O assunto acaba por requerer uma reflexão mais cautelosa que poderá até se consubstanciar em uma proposta de sistematização do uso da mineração de textos como ferramenta de apoio à indexação visando o aumento do índice de precisão de resposta nos sistemas de recuperação da informação, mas comprovadamente, não poderá exercer o papel que a indexação manual desempenha nos sistemas de informação. A postura racional está em adotar a mineração de textos e a indexação manual como recursos complementares.

A primeira tese defendida impacta o cumprimento do primeiro objetivo específico do trabalho que é o de avaliar se a recuperação por intermédio da ferramenta de mineração de textos traz ganho no índice de precisão quando comparada à lista de palavras-chave utilizadas na indexação manual por bibliotecários da Caixa Econômica Federal na base de dados do Infohab. O ganho de precisão é a medida que permite inferir que a mineração de textos não pode substituir a indexação manual, fato explícito na comprovação do primeiro pressuposto da investigação.

A comprovação do primeiro pressuposto abre caminho para a possibilidade de utilização das potencialidades da ferramenta de mineração de textos de modo combinado com o processo de indexação, fato verificado diante dos resultados obtidos com a metodologia empregada. Daí a defesa da primeira tese.

Nas teses subseqüentes procura-se entender o papel da mineração de textos no processo de indexação.

Tese II - A mineração de textos pode ser considerada ferramenta de indexação automática por extração automática de termos, desde que seja incluído neste processo o julgamento dos indexadores que deverão selecionar os termos a serem usados na representação do conteúdo dos documentos.

A segunda tese defendida está relacionada com uma potencialidade da ferramenta de mineração de textos que é reforçada pela comprovação do segundo e terceiro pressupostos da pesquisa.

Nestes pressupostos, em todas as perguntas formuladas ao protótipo com mineração de textos em comparação com a lista de palavras utilizadas na indexação manual, obtém-se como resposta uma quantidade maior de itens bibliográficos e resultados sempre maiores do que zero, ou seja, sempre são recuperados itens bibliográficos no protótipo mesmo quando o resultado com a lista de palavras-chave utilizadas na indexação manual é nulo, independentemente do cálculo do índice de precisão. Isto equivale dizer que o índice de precisão poderá ser baixo, mas a recuperação de itens bibliográficos será sempre realizada em quantidades razoáveis.

O protótipo com uso da ferramenta de mineração de textos (*Software* BR/*Search*) constrói automaticamente as listas por dois tipos de descobertas: por listas de conceitos-chave, usando as palavras mais freqüentes que ocorrem em cada texto de cada item bibliográfico recuperado e por descrição de classes de textos, realizando a mesma operação só que com o resultado total da pesquisa. Esta possibilidade, somada à recuperação de qualquer termo ou caractere contido no texto, confere à ferramenta sua grande capacidade de recuperação e montagem de listas de palavras mais freqüentes, bem como análise automática de freqüência de palavras por item bibliográfico ou pelo resultado total da pesquisa.

As listas de palavras mais freqüentes e a análise automática de freqüência de palavras são obtidas por extração automática de termos, desde que o trabalho do indexador no julgamento dos termos extraídos seja decisivo na escolha dos termos que deverão ser usados para indexar os documentos, conforme colocação da página 175 deste trabalho. Esta condição é essencial para a proposta de montagem de um sistemática de uso dos termos gerados a partir da mineração de textos apoiando o processo de indexação no aumento do índice de precisão de resposta no processo de busca e recuperação da informação.

A segunda tese também impacta o cumprimento do segundo objetivo da investigação, ou seja, esquadrinhar o uso dos termos resultantes do emprego de ferramenta de mineração de textos no enriquecimento da lista de palavras-chave utilizada no Infohab, a fim de aprimorar a indexação manual em relação ao índice de precisão de resposta na recuperação da informação. Este objetivo não só foi cumprido como tornou-se um mecanismo importante junto com a verificação dos pressupostos 1 e 2, na proposição de uma sistemática que combine o uso da mineração de textos com a indexação manual para o aprimoramento do processo de indexação e da conseqüente recuperação da informação.

A tese em questão enseja ainda que as listas de palavras mais freqüentes dos textos, convertidas pela ação do indexador em um núcleo de palavras que poderão ser empregadas para representar o conteúdo dos documentos, possam ser a base para definir o papel da mineração de textos no processo de indexação, eliminando o erro de considerar que a substituição da indexação manual por uma ferramenta de mineração de textos que possa, por extração automática de termos, ser a solução ideal.

Tese III - A mineração de textos, desde que entendida como indexação automática, pode ser considerada um instrumento de apoio na construção e/ou enriquecimento do vocabulário controlado, que é gerado e utilizado na indexação manual.

A terceira tese é o complemento da segunda na definição do papel da mineração de textos no processo de indexação, além de arrematar a proposta da sistemática que combina mineração de textos com indexação manual. Se é possível considerar a ferramenta de mineração de textos como indexação automática, o que é factível pelos resultados até aqui alcançados, ela poderá ser usada como efetivo instrumento de enriquecimento de uma lista de palavras-chave ou até mesmo de um tesauro.

A terceira tese reforça que a indexação "automática" por extração automática de termos, deve ser encarada como instrumento complementar à indexação manual. Isto equivale dizer que ferramentas de mineração de textos devem ser utilizadas, no âmbito da indexação, com vistas à melhoria contínua do índice de precisão da resposta no processo de busca e recuperação da informação, como elementos de apoio às decisões do indexador, buscando tornar a indexação o menos possível subjetiva em sua tarefa de prescrever termos para representar o conteúdo dos documentos.

Os resultados alcançados com as indagações propostas pelo estudo permitem afirmar que as três teses defendidas são absolutamente factíveis, não só pelo verificação dos pressupostos, mas por apoiar a construção da contribuição que se pretende efetiva na construção do conhecimento na Ciência da Informação. Por isso, a proposta de junção de ferramentas de mineração de dados, hoje cada vez mais difundidas como "tábua de salvação" para a análise de grandes quantidades de dados e textos em grandes *data warehouse* corporativos[19], com processos tradicionais consolidados e testados como é o caso do processo de indexação. Nesta problemática, certamente está uma parte da discussão da empregabilidade das tecnologias da informação na melhoria dos processo de busca e recuperação de informação útil às demandas dos usuários. Este é o desafio e, provavelmente, o paradigma dos modernos sistemas de informação: o atendimento efetivo e proativo das demandas de usuários e corporações.

[19] ***Data warehouse*** **corporativo** – coleção de banco de dados orientados e integrados ao assunto designado a suportar a função de sistemas de apoio à decisão no âmbito de uma corporação. Um *data warehouse* corporativo retém a maioria dos dados atômicos que a corporação possui. Dois ou mais *data warehouses* corporativos podem ser combinados para criar um *data warehouse* distribuído (Inmon & Hackathorn, 1994).

9- CONCLUSÕES

Durante a elaboração da presente pesquisa algumas conclusões preliminares apoiaram o seu desenvolvimento, contribuindo assim, para o alcance dos objetivos propostos:

- O aperfeiçoamento contínuo do processo de indexação deve passar pelo conhecimento proativo das necessidades dos usuários, o que deve proporcionar subsídios para a determinação dos requisitos a serem utilizados no âmbito do gerenciamento estratégico da informação;

- Em termos práticos, a mineração de textos, no âmbito do processo de busca e recuperação da informação, poderá ser pensada como uma ferramenta a ser empregada na busca da melhoria das respostas nestes sistemas. A avaliação desta possibilidade poderá ser materializada a partir da utilização do índice de precisão para aferir a performance da ferramenta nesta tarefa, bem como compará-la aos instrumentos tradicionais, utilizados hoje;

- A utilização dos índices de precisão deverá sempre propiciar parâmetros para a melhoria contínua da resposta obtida dos sistemas de recuperação da informação. O conhecimento das necessidades de informação dos usuários é ponto de partida para a concretização desta meta; entretanto, tais necessidades acabam por gerar também graus de imprecisão, tal como observa Foskett (1996), ou seja, incapacidade de um sistema de informação de recuperar documentos úteis frente à solicitação do usuário, sobretudo se envolver a negociação da questão;

- Os autores estudados concordam com o papel preponderante que cabe ao julgamento dos usuários para o cálculo do índice de precisão. Com este dado torna-se factível pensar na precisão como um elemento

importante de análise e decisão na busca da melhor resposta nos sistemas de busca e recuperação da informação. Não é possível calcular o número de documentos úteis encontrados pelo sistema sobre o total de documentos relevantes contidos no sistema, sem o julgamento do usuário que demanda tais documentos, ou seja, a precisão se consubstancia por meio de um julgamento externo ao sistema de busca e recuperação da informação, o que vai determinar também, a sua capacidade de atendimento ou o seu desempenho; e

- A precisão não se dá *per se*, mas no contexto em que operam a revocação, a exaustividade a especificidade e, sobretudo, tendo como ponto de equilíbrio, o usuário que vai definir, em nome da sua necessidade de informação, o que é útil ou inútil dentre toda a informação recuperada.

Com base na proposta do estudo comparado entre a mineração de textos e a indexação manual, a investigação concentrou-se na avaliação da resposta obtida no processo de busca e recuperação da informação, por meio de uma medida objetiva, o índice de precisão. Desde a década de 70, a questão da precisão foi amplamente discutida em associação com a análise de desempenho de um sistema de recuperação da informação. Para Lancaster & Fayen (1973), ao considerar os fatores que interferem no desempenho destes sistemas, será necessário conhecer anteriormente os pré-requisitos do usuário em relação aos resultados de busca e recuperação da informação.

A escolha do índice de precisão permitiu avaliar, em termos percentuais, o desempenho de um protótipo com aplicação de mineração de textos, confeccionado para ser o espelho da amostra selecionada da Base do Infohab, onde os documentos são indexados manualmente. O resultado da avaliação dos desempenhos do protótipo e da Base do Infohab atingiu um dos objetivos do estudo, qual seja, avaliar se a recuperação da informação por meio de ferramenta

de mineração de textos traz ganho no índice de precisão se comparada à lista de palavras-chave utilizadas na indexação manual na Base de dados do Infohab.

A avaliação do desempenho do protótipo permitiu também a verificação da viabilidade de se utilizar os termos resultantes do emprego da ferramenta de mineração de textos no enriquecimento da lista de palavras-chave utilizadas na indexação manual. Apesar da ferramenta constituir um importante instrumento na identificação de palavras-chave, o indexador continua como um dos principais artífices no processo de indexação, dada a sua competência na escolha de quais termos serão usados para identificar o conteúdo dos documentos.

Este debate, apesar de já antigo sempre retorna como nova proposição a cada desenvolvimento de novas tecnologias, trazendo expectativas para a resolução dos problemas relacionados ao processo de busca e recuperação da informação. Para Lancaster (1993), entretanto, apesar de terem ocorrido muitos avanços no processamento da linguagem natural por computador, é mister admitir que a "compreensão" de textos pelo computador ainda se acha muito limitada. Ou seja, é possível construir instrumentos auxiliares morfológicos, sintáticos e semânticos que ajudem o computador a interpretar textos, mas isto ainda está muito longe do que acontece quando um ser humano lê um texto e compreende o que o autor quer dizer.

A mineração de textos apresenta-se, neste contexto, como uma ferramenta de ponta, cujo objetivo, segundo Feldman & Hirsh *apud* Wives (1999), é constituir-se em um meio efetivo de recuperação, filtragem, manipulação e resumo do conhecimento contido em grandes volumes de informações textuais, para apresentá-lo em forma de gráficos, listas ou tabelas.

Esta possibilidade da ferramenta de mineração de textos, abriu caminho para a proposição de uma sistemática (disposta na figura 31 da página 184) de utilização dos termos gerados a partir da mineração de textos que apóia o processo de indexação manual visando o aumento do índice de precisão de resposta no processo de busca e recuperação da informação, alcançando assim, o terceiro e último objetivo proposto.

De acordo com McGarry (1999), os sistemas de computador obedecem a algoritmos, mas o conteúdo semântico dos textos está além de sua compreensão. As mentes humanas têm conteúdos semânticos, significados e ressonâncias, o que enseja a discussão fomentada pelo autor a respeito do futuro da Ciência da Informação: poderá esta Ciência ser conduzida dentro de um sistema fechado de raciocínio algorítmico? Poderá crescer e desenvolver-se no vazio cultural?

Apesar de toda a discussão em torno da utilização de um ou outro método, ou seja, a indexação manual ou a mineração de textos, os resultados, na verdade, só poderão ser validados por intermédio da avaliação dos usuários. Isto significa dizer que os sistemas de recuperação da informação, além de buscar atender às demandas informacionais dos usuários, dependem destes para que a qualidade dos seus serviços seja reconhecida.

10- CONTRIBUIÇÃO E LIMITAÇÕES DO ESTUDO

A partir do estudo realizado até aqui pode-se considerar como contribuições da pesquisa:

- Fomento do uso de tecnologias da informação, notadamente a mineração de textos na descoberta de conhecimento em bases textuais no aprimoramento de processos tradicionais da Ciência da Informação. Com isso, abre-se um campo profícuo na modernização e adaptação de novas soluções na busca da melhoria do desempenho dos sistemas de informação;

- Associação do processo de indexação à mineração de textos, visando a melhoria do índice de precisão na busca e recuperação da informação, onde o usuário continua a ter um papel preponderante no julgamento da utilidade dos resultados obtidos com as bases de dados, daí o uso da medida objetiva do índice de precisão que o envolve na validação da utilidade da informação recuperada. Neste sentido, o estudo procura unir em uma mesma sistemática, o processo de indexação, a mineração de textos e o julgamento do usuário na melhoria da resposta no PBRI;

- Desenvolvimento de uso da ferramenta de mineração de textos aplicada aos processos de recuperação da informação, já que foi percebido no estudo da literatura na área da Ciência da Informação e da Ciência da Computação, uma lacuna quanto à esta possibilidade; e

- Ampliação da perspectiva de emprego da indexação combinada à tecnologias da informação na gestão da precisão, na recuperação de informação em bases de dados corporativas, possibilitando a descoberta de conhecimento que pode ser usado no processo decisório, bem como no desenvolvimento de novos empreendimentos.

A pesquisa apresenta, também, limitações que basicamente podem ser colocadas em duas observações:

- Literatura incipiente sobre o assunto tanto na área de domínio da Ciência da Informação como na quase inexistência de trabalhos publicados no periódico Ciência da informação. Neste último caso, nenhum trabalho no formato de tese ou dissertação foi encontrado nas bases de dados brasileiras. Esta escassez traz, como conseqüência imediata, a constatação de que a mineração de textos é um campo aberto inclusive na formação de uma terminologia própria que hoje é, em sua totalidade, adaptada da mineração de dados, área de pesquisa mais consolidada como constata um dos autores citados na investigação (Benoît, 2002); e

- O estudo de um único caso, o do Infohab. A proposição de novos estudos de caso trarão novos relatos de experiência, contribuindo para a efetiva construção da terminologia e das experiências de junção entre a mineração de textos, o processo de indexação, o índice de precisão e o processo de busca e recuperação da informação. A construção conjunta do conhecimento entre estas áreas de pesquisa, que funde interesses da Ciência da Informação com a Ciência da Computação, dependerá de novos trabalhos com novos casos e experiências.

A escassez de literatura e estudos práticos sobre o tema é uma boa oportunidade para os pesquisadores buscarem consenso quanto ao uso de definições e termos mais precisos, fato que ainda não ocorreu. Esta situação também permite sugerir, com maior flexibilidade, temas para pesquisas vindouras.

11- SUGESTÕES PARA NOVAS PESQUISAS

Todo empreendimento de pesquisa tem embutido nas suas conclusões o ponto de partida para novos estudos e investigações, pois a solução dos problemas levantados, invariavelmente leva a formular novos pressupostos em um ciclo de construção e consolidação do conhecimento.

Assim, esta pesquisa poderia ser o ponto de partida para:

- Estudos sobre a possibilidade de uso da mineração de textos para determinar o grau de equilíbrio entre a revocação e a especificidade no processo de busca e recuperação da informação;

- Estudos sobre a possibilidade de aplicação da mineração de textos como ferramenta de avaliação e controle da coerência da indexação no aspecto de quantidade de termos atribuídos;

- Estudos sobre a proposição de um plano diretor de sistemas para o desenvolvimento de *software* de mineração de textos, que inclua o requisito de empregabilidade nos sistemas de recuperação da informação, bem como o desenvolvimento de interfaces amigáveis para o usuário e o profissional da informação.

- Diagnósticos estratégicos a serem realizados nos sistemas de informação, a fim de verificar se há de fato a utilização sistemática de instrumentos de avaliação da precisão de resposta no processo de busca e recuperação da informação, com vistas ao aprimoramento contínuo do seu desempenho;

- Estudos de modelos de avaliação sistemática do índice de precisão de resposta e/ou de revocação no processo de busca e recuperação da informação nos sistemas de apoio à decisão;

12- REFERÊNCIAS

ALCAIDE, G. S. *et al.* Análise comparativa e de consistência entre representações automática e manual de informações documentárias. *Transinformação*, jan./jun.,v.13, n.1, p.23-41, 2001.

AMARAL, F. C. N. do. *Data mining: técnicas e aplicações para o marketing direto*. São Paulo : Berkeley, 2001.

ANDERSON, J. D. & PÉREZ-CARBALLO, J. The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: research, and the nature of human indexing. *Information Processing and Management*, March, v.37, n.2, p. 231-254, 2001.

ANDERSON, J. D. & PÉREZ-CARBALLO, J. The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part II: machine indexing, and the allocation of human versus machine effort. *Information Processing and Management*, March, v.37, n.2, p. 255-277, 2001.

ARAUJO, V. M. R. H. de. *Sistemas de recuperação da informação: nova abordagem teórico-conceitual*. Rio de Janeiro : Universidade Federal do Rio de Janeiro, 1994. (tese de doutorado em Ciência da Informação).

ARAÚJO JR., R. H. de. *Estudo de necessidades de informação dos gerentes do setor editorial e gráfico do Distrito Federal*. Brasília : Universidade de Brasília, 1998. (dissertação de mestrado em Ciência da Informação).

BARANOW, U. G. In: Conferência Brasileira de Classificação Bibliográfica. Rio de Janeiro, 12 a 17 de setembro de 1976. Anais. Rio de Janeiro : IBICT; Brasília : ABDF. *Aspectos lingüísticos de linguagens de indexação*, p. 295-310.

BARANOW, U. G. Perspectivas na contribuição da lingüística e de áreas afins à Ciência da Informação. *Ciência da Informação*, v.12, n.1, p. 23-35, jan./abril, 1983.

BASTOS, S. B. *Análise comparativa entre indexação automática e manual da literatura brasileira de ciência da informação*. Brasília : Universidade de Brasília, 1984 (dissertação de mestrado em Biblioteconomia e Documentação).

BELKIN, N. J. In: JONES, K. S. *Information retrieval experiment*. London : Butterworths, 1981. *Ineffable concepts in information retrieval*, p. 44-58.

BELKIN, N. J. & CROFT, W. B. Retrieval techniques. *Annual Review of Information Science and Technology*, v.22, p. 41-76, 1987.

BENOîT, G. Data mining. *Annual Review of Information Science and Technology*, v. 36, p. 265-310, 2002.

BORKO, H. & BERNIER, C. L. *Indexing concepts and methods*. New York : Academic Press, 1978.

CABENA, P. *et al. Discovering data mining: from concept to implementation*. New Jersey : Prentice Hall, 1997.

CAMPOS, V. F. *TQC: controle da qualidade total: no estilo japonês*. Belo Horizonte : Fundação Christiano Ottoni : Escola de Engenharia da UFMG, 1992.

CARO, C.; CEDEIRA, L. & TRAVIESO, C. La investigación sobre recuperación de información desde la perspectiva centrada en el usuario: métodos y variables. *Revista Española de Documentación Científica*, v.26, n.1, p. 40-55, enero/marzo, 2003.

CENTRO DE REFERENCIA E INFORMAÇÃO EM HABITAÇÃO – INFOHAB. *Regimento – Infohab*. Niterói : Infohab, 2000.

CHECKLAND, P. *Systems thinking, systems practice*. Chichester : John Wiley & Sons, 1999.

CHOMSKY, N. *Linguagem e mente: pensamentos atuais sobre antigos problemas*. Tradução de Lúcia Lobato. Brasília : Editora Universidade de Brasília, 1988.

CIANCONI, R. de B. Sistemas de recuperação em linha. *Ciência da Informação*, v.19, n.1, p. 131-136, jul./dez., 1990.

CINTRA, A. M. M. Elementos de lingüística para estudos de indexação. *Ciência da Informação*, v.12, n.1, p. 5-22, jan./abril, 1983.

CINTRA, A. M. M. *Para entender as linguagens documentárias*. São Paulo : Polis, 1994. (Coleção Palavra Chave, 4).

CHUNG, W. *The automatic text mining framework for knowledge discovery on the Web*. Arizona : University of Arizona, 2004. (tese de doutorado em Ciência da Computação).

CLEVERDON, C. W. *Report on testing and analysis of investigation into comparative efficiency of indexing systems*. Cranfield : Aslib, 1962.

CLEVERDON, C. W. *The testing of index language devices*. Aslib Proceedings, n.15, p.106-130, 1963.

COLE, C. Intelligent information retrieval: part IV. Testing the timing of two information retrieval devices in a naturalistic setting. *Information Processing and Management*, v.37, n.1, p. 163-182, 2001.

COSTA, J. A. & MELO, A. S. e. *Dicionário da língua portuguesa*. Porto : Porto Editora, 1989.

CUNHA, I. M. R. F. *Do mito à análise documentária*. São Paulo : Edusp, 1990.

DAGHLIAN, J. *Lógica e álgebra de Boole*. São Paulo : Atlas, 1995.

DAVIES, R. The creation of new knowledge by information retrieval and classification. *Journal of Documentation*, v.45, n.4, December, 1989.

DELMATER, R. & HANCOCK, M. *Data mining explained: a manager's guide to customer-centric business intelligence*. Woburn : Butterworth-Heinemann, 2001.

DIAZ, J. E. B. *Além dos meios e mensagens: introdução à comunicação como processo, tecnologia, sistema e ciência*. Petrópolis : Vozes, 1986.

DIXON, M. *An overview of document mining technology*. Available at: http://www.software.ibm.com. Acesso em: 20/09/2004.

DUNLOP, M. Reflections on Mira: interactive evaluation in information retrieval. *Journal of the American Society for Information Science*, v.51, n.14, p. 1269-1274, 2000.

FELDMAN, R. & HIRSH, H. Exploiting background information in knowledge discovery from text. *Journal of Intelligent Information Systems*, v.9, n.1, p.83-97, July/August, 1997.

FERREIRA, A. B. de H. *Novo dicionário da língua portuguesa*. Rio de Janeiro : Nova Fronteira, 1988.

FOSKETT, A. C. *The subject approach to information*. 5th ed. London : Unipub, 1996.

FRASER, P. *Pistes d'exploration pour l'elaboration d'un système formel de montée en abstraction et d'emergence de categorisations*. Canadá : Université Laval, 2001. (dissertação de mestrado em Ciência da Computação).

FROEHLICH, T. J. Relevance reconsidered-towards an agenda for 21st century. *Journal of the American Society for Information Science*, v.45, n.3, p. 124-134, 1994.

GARDIN, J. C. *La logique du plausible*. Paris : Maison des sciences de l'homme, 1981.

GARVIN, P. L. *Natural language and computer*. New York : McGraw-Hill, 1963.

GIRJU, C. R. *Text mining is a rapidly emerging field concerned with the extraction of concepts, relations and implicit knowledge*. Dallas : University of Texas, 2002. (tese de doutorado em Ciência da Computação).

GOLDMAN, J. A. *A digital model for data mining of text documents: databases, computacional linguistics*. Los Angeles : University of California, 1998. (tese de doutorado em Ciência da Computação).

GROGAN, D. *A prática do serviço de referência*. Tradução de Antonio Agenor Briquet de Lemos. Brasília : Briquet de Lemos/Livros, 1995.

GUIMARÃES, J. A. C. In: RODRIGUES, G. M. & LOPES, I. L. (org.) Organização e representação do conhecimento na perspectiva da Ciência da Informação. Brasília : Thesaurus, 2003. (Série: Estudos Avançados em Ciência da Informação, v.15). *A*

análise documentária no âmbito do tratamento da informação: elementos históricos e conceituais, 100-117.

HALAL, E. & KULL, M. *Measuring organizational intelligence*. Disponível em: http://www.auburn.edu/administration/horizon/measuring.html. Acesso em: 12/02/2004.

HOOPER, R. S. *Indexer consistency tests: origin, measurements, results and utilization*. Bethesda : IBM, 1965.

INGWERSEN, P. Cognitive information retrieval. *Annual Review of Information Science and Technology*, v.34, p. 3-52, 1999.

INMON, W. H. & HACKATHORN, R. D. *Using the data warehouse*. New York : John Wiley & Sons, 1994.

INMON, W. H.; TERDEMAN, R. H. & IMHOFF, C. *Exploration warehousing: turning business information into business opportunity*. New York : John Willey & Sons, 2000.

KERLINGER, F. N. *Metodologia da pesquisa em ciências sociais: um tratamento conceitual*. Tradução de Helena Mendes Rotundo. EPU : São Paulo, 1980

KOBLAS, J. Beyond information quality: fitness for purpose and electronic information resource use. *Journal of Information Science*, v. 21, n.2, p.95-114, 1995.

KUHLTHAU, C. C. Inside the search process: information seeking from the user's perspective. *Journal of the American Society for Information Science*, v.42, p. 361-371, 1991.

KUHLTHAU, C. C. *Seeking meaning: a process approach to library and information services.* Norwood : Ablex, 1993.

KUHLTHAU, C. C. Accommodating the user's information search process: challenges for information retrieval system designers. *Bulletin of the American Society for Information Science*, v.25, n.3, p. 12-16, 1999.

LANCASTER, F. W. *Indexação e resumos: teoria e prática.* Tradução de Antonio Agenor Briquet de Lemos. Brasília : Briquet de Lemos/Livros, 1993.

LANCASTER, F. W. *Indexing and abstracting in theory and practice.* 2th edition. London : Library Association, 1998.

LANCASTER, F. W. *Information retrieval systems: characteristics, testing and evaluation.* 2th edition. New York : Wiley-Interscience, 1979.

LANCASTER, F. W. & FAYEN, E. G. *Information retrieval on-line.* Los Angeles : Melville, 1973.

LANCASTER, F. W. & WARNER, A. J. *Information retrieval today.* Arlington : Information Resourses Press, 1993.

LE COADIC, Y. F. *A ciência da informação.* Tradução de Maria Yêda F. S. de Filgueiras Gomes. Brasília : Briquet de Lemos/Livros, 1996.

LE COADIC, Y. F. *La science de l'information.* Paris : Presses Universitaires de France, 1994.

LOPES, I. L. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. *Ciência da Informação*, v.31, n.1, p. 41-52, jan./abr., 2002.

MAGALHÃES, M. N. & LIMA, A. C. P. *Noções de probabilidade e estatística*. São Paulo : IME-USP, 2001.

MAIRI, H. Dos fundamentos da significação à produção do sentido. *Perspectivas em Ciência da Informação*, v.1, n.1, p.93-109, jan./jun., 1996.

McGARRY, K. *O contexto dinâmico da informação: uma análise introdutória*. Tradução de Helena Vilar de Lemos. Briquet de Lemos/Livros, 1999.

McGEE, J. & PRUSAK, L. *Gerenciamento estratégico da informação: aumente a competitividade e a eficiência de sua empresa utilizando a informação como ferramenta estratégica*. Tradução de Astrid B. de Figueiredo. Rio de Janeiro : Campus, 1994.

MEIDEIROS, M. B. B. *Tratamento automático de ambigüidades na recuperação da informação*. Brasília : Universidade de Brasília, 1999. (tese de doutorado em Ciência da Informação).

MIKE, S. et al. In: Special Interest Group on Information Retrieval, SIGIR, VII. Proceedings. London : Springer-Verlag, 1994. *A full-text retrieval system with a dynamic abstract generation function*.

MORRIS, J. & HIRST, G. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, v.17, n.1, March, 1991.

MOSCAROLA, J. *et al. Technology watch via textual data analysis*. France : Université de Savoie, 1998.

NORTON, M. J. Knowledge Discovery in Databases. *Library Trends,* v.48, n.1, p.9-21, Summer 1999.

NOVELLINO, M. S. F. A linguagem como meio de representação ou de comunicação da informação. *Perspectivas em Ciência da Informação*, v.3, n.1, p.137-146, jul./dez., 1998.

O'BRIEN, J. A. *Introduction to information systems*. 9th edition. New York : Irwin McGraw-Hill, 2000.

ONG, T. H. *Language – and domain – independent knowledge maps: a statistical phrase indexing approach*. Arizona : University of Arizona, 2004. (tese de doutorado em Ciência da Computação).

ORRICO, E. G. D. *Binômio lingüística – ciência da informação: abordagem teórica para elaboração de metafiltro de recuperação da informação*. Rio de Janeiro : Universidade Federal do Rio de Janeiro, 2001. (tese de doutorado em Ciência da Informação).

PEREIRA, V. L. C. *Sistemas de redução da informação: uma (ir) recuperação metodologicamente configurada*. Rio de Janeiro : Universidade Federal do Rio de Janeiro, 1994. (dissertação de mestrado em Ciência da Informação).

POINCARÉ, J. H. *A ciência e a hipótese*. Tradução de Maria Auxiliadora Kneipp. Brasília : Editora Universidade de Brasília, 1988.

POLANCO, X. & FRANÇOIS, C. In: Proceedings of the Sixth International ISKO Conference, Toronto, 10-13 July, 2000. Ergon Verlag : Würzburg, 2000. *Data clustering and cluster mapping or visualization in text processing and mining*, p.359-365.

POUNTAIN, D. Sorting out the sorts. *Byte*, v.12, p. 275-280, 1987.

RAMOS, M. G. *O uso da teoria de função de crença em sistemas de recuperação da informação*. Brasília : Universidade de Brasília, 1999. (dissertação de mestrado em Ciência da Informação).

ROBERTSON, S. E. In: JONES, K. S. *Information retrieval experiment*. London : Butterworths, 1981. *The methodology of information retrieval experiment*, p. 9-31.

ROBREDO, J. & CUNHA, M. B. da. *Documentação de hoje e de amanhã*. Brasília : ed. autor, 1986.

ROBREDO, J. *Da ciência da informação revisitada aos sistemas humanos de informação*. Brasília : Thesaurus; SSRR Informações, 2003.

ROCKART, J. F. Chief executives define their own data needs. *Harvard Business Review*, v. 57, n. 2, p. 81-93, March/April, 1979.

ROWLEY, J. *A biblioteca eletrônica*. Tradução de Antonio Agenor Briquet de Lemos. Brasília : Briquet de Lemos/Livros, 2002.

ROWLEY, J. *Abstracting and indexing*. 2th edition. London : Clive Bingley, 1988.

RUIZ, R. P. *El analisis documental: bases terminológicas, conceptualización y estructura operativa*. Granada : Universidad de Granada, 1992.

SARACEVIC, T. Information science. *Journal of the American Society for Information Science*, v.50, n.12, p. 1051-1063, 1999.

SARACEVIC, T. Interdisciplinary nature of information science. *Ciência da Informação*, v.24, n.1, p. 36-41, jan./abril, 1995.

SARACEVIC, T. In: SARACEVIC, T. *Introduction to information science*. New York : R. R. Bowker, 1970. *The concept of relevance in information science: a historical review*, p. 111-151.

SARACEVIC, T. In: *Proceedings of the 18th International conference on research and development in information retrieval* Seattle : Association of Computing Machinery, 1995. *Evaluation of evaluation in information retrieval*, p. 138-146.

SILVA, E. M. *Descoberta de conhecimento com uso de text mining: cruzando o abismo de Moore*. Brasília : Universidade Católica de Brasília, 2002. (dissertação de mestrado em Ciência da Computação).

SLAMECKA, V. Methods and research for design of information networks. *Library Trends*, v.18, p. 551-568, 1970.

SLAMECKA, V. & JACOBY, J. *Effect of indexing aids on the reliability of indexers: final technical note*. Bethesda : Documentation Inc., 1963.

SLYPE, G. V. *Los lenguajes de indización: concepción, construcción y utilización en los sistemas documentales*. Madrid : Fundación Germán Sánchez Ruipérez, 1991.

SWIFT, R. *Accelerating customer relationship: using CRM and relationship techonologies*. New Jersey : Prentice Hall, 2000.

TÁLAMO, M. de F. G. M.; LARA, M. L. G. de & KOBASHI, N. Y. Contribuição da terminologia para a elaboração de tesauros. *Ciência da Informação*, v.21, n.3, p.197-200, set./dez. 1992.

TARAPANOFF, K. (Org.) *Inteligência organizacional e competitiva*. Brasília : Editora Universidade de Brasília, 2001.

TARAPANOFF, K.; ARAÚJO JR., R. H. de & CORMIER, P. M. J. Sociedade da informação e inteligência em unidades de informação. *Ciência da Informação*, v.29, n.3, p.91-100, set./dez. 2000.

TARAPANOFF, K.; QUONIAM, L.; ARAÚJO JR., R. H. de & ALVARES, L. Intelligence obtained by applying data mining to a database of French theses on the subject of Brazil. *Information Research*, v. 07, n. 01, October, 2001. Available at: http://InformationR.net/ir/7-1/paper117.html. Acesso em: 20/09/2004.

TAYLOR, R. S. *Value-added process in information systems*. Norwood : Ablex, 1986.

TINKER, J. F. Imprecision in indexing. *American Documentation*, v.17, p.93-103, 1966; v.19, p.322-330, 1968.

TRYBULA, W. J. Text mining. *Annual Review of Information Science and Technology*, v.34, p. 385-419, 1999.

TRYBULA, W. J. *Text mining and knowledge discernment: an exploratory investigation*. Austin : University of Texas, 1999. (tese de doutorado em Ciência da Computação).

TURBAN, E.; McLEAN, E. & WETHERBE, J. *Tecnologia da informação para gestão*. Porto Alegre : Bookman, 2004.

VIZCAYA, D. A. Lenguaje e información. *Data Grama Zero – Revista de Ciência da Informação*, v.2, n.4, ago., 2001. Disponível em: http://www.dgzero.org. Acesso em: 30/01/2005.

WELLISCH, H. H. *Indexing from A to Z*. 2nd ed. New York : H. W. Wilson, 1995.

WILSON, T. D. Models in information behaviour research. *Journal of Documentation*, v.55, n.3, p. 249-270, June, 1999.

WISTON, P. *Rethinking artificial intelligence*. Massachusetts : MIT, 1997.

WIVES, L. K. *Estudo sobre agrupamento de documentos textuais em processamento de informação não estruturadas usando técnicas de clustering.* Porto Alegre : Universidade Federal do Rio Grande do Sul, 1999. (dissertação de mestrado em Ciência da Computação).

WIVES, L. K. & LOH, S. *Tecnologias de descoberta de conhecimento em informações textuais: ênfase em agrupamento de informações*. Porto Alegre : PPGC/Universidade Federal do Rio Grande do Sul, 2000.

ZAVITOSKI, M. T. *Exploração do uso do tesauro como instrumento de recuperação da informação*. São Paulo : Universidade de São Paulo, 2001. (dissertação de mestrado).

ANEXOS

ANEXO 1 –Formulários de coleta de dados

1	UNIVERSIDADE DE BRASÍLIA - UnB Faculdade de Economia, Administração, Contabilidade e Ciência da Informação – FACE Departamento de Ciência da Informação e Documentação - CID Programa de Pós-Graduação em Ciência da Informação
	Pesquisa Bibliográfica na Base de Dados do Centro de Referência e Informação em Habitação - INFOHAB

NOME		UNIDADE
FUNÇÃO	TELEFONE	E-MAIL

DADOS RELATIVOS À PESQUISA (descreva em texto livre sua necessidade de informação)

PALAVRAS-CHAVE (descreva a forma como a pesquisa foi submetida ao sistema: termos usados e outras especificações: operadores, busca em campos, etc.)

FINALIDADE DA PESQUISA
☐ Dissertação/Tese ☐ Monografia ☐ Pesquisa ☐ Outros _____

RESULTADO DA PESQUISA (nº total de documentos encontrados pelo sistema)
Obs.: anexar o resultado da pesquisa caso necessário

VALIDAÇÃO DO USÚARIO (nº de documentos úteis recuperados pelo sistema)
Obs.: anexar o resultado da pesquisa caso necessário

CÁLCULO DA PRECISÃO DA PESQUISA (%)
Precisão = $\frac{\text{Número de documentos úteis recuperados pelo sistema}}{\text{Número total de documentos encontrados pelo sistema}}$ X 100

PERÍODO DA PESQUISA
DATA: ——— / ——— / ———.

2	**UNIVERSIDADE DE BRASÍLIA - UnB** **Faculdade de Economia, Administração, Contabilidade e Ciência da Informação – FACE** **Departamento de Ciência da Informação e Documentação - CID** **Programa de Pós-Graduação em Ciência da Informação**
	Pesquisa Bibliográfica na Base de Dados do Centro de Referência e Informação em Habitação – INFOHAB *(Protótipo com Aplicação da Mineração de Textos)*

NOME		UNIDADE
FUNÇÃO	TELEFONE	E-MAIL

DADOS RELATIVOS À PESQUISA (descreva em texto livre sua necessidade de informação)

PALAVRAS-CHAVE (descreva a forma como a pesquisa foi submetida ao sistema: termos usados e outras especificações: operadores, busca em campos, etc.)

FINALIDADE DA PESQUISA
☐ Dissertação/Tese ☐ Monografia ☐ Pesquisa ☐ Outros _____

RESULTADO DA PESQUISA (n° total de documentos encontrados pelo sistema)
Obs.: anexar o resultado da pesquisa caso necessário

VALIDAÇÃO DO USÚARIO (n° de documentos úteis recuperados pelo sistema)
Obs.: anexar o resultado da pesquisa caso necessário

CÁLCULO DA PRECISÃO DA PESQUISA (%)
$\text{Precisão} = \frac{\text{Número de documentos úteis recuperados pelo sistema}}{\text{Número total de documentos encontrados pelo sistema}} \times 100$

PERÍODO DA PESQUISA
DATA: ——— / ——— / ———.

ANEXO 2 – Lista dos itens bibliográficos constantes da amostra da pesquisa

1. CAVALCANTI, J. C. S. *Setor de saneamento no brasil: estrutura, dinâmica e perspectivas*. Rio de Janeiro : UFRJ, 1987. 214p. Registro no Infohab: nº 72023.

2. BONDAROVSKY,S. H. *Regulamentação para uso racional da água*. Rio de Janeiro : UFRJ, 2004. Registro no Infohab: nº 92810.

3. AZEVEDO, J. *Estudo ambiental e econômico do composto orgânico do sistema de beneficiamento de resíduos sólidos urbanos da usina de Irajá, município do Rio de Janeiro*. Rio de Janeiro : sem editora, 2000. Registro no Infohab: nº 86242.

4. AZEVEDO, J.; SILVA FILHO, V. & DAMASCENO, R. N. *et al.* Panorama das usinas de beneficiamento de resíduos sólidos urbanos do estado do Rio de Janeiro. Rio de Janeiro : sem editora, 2000. Registro no Infohab: nº 87715.

5. AZEVEDO, J.; SILVA FILHO, V. & DAMASCENO, R. N *et al. Valor agrícola e comercial do composto orgânico de resíduos sólidos urbanos da usina de Irajá, município do Rio de Janeiro*. Porto Alegre :PUC, 2001. Registro no Infohab: nº 87714.

6. XAVIER, S. M. *Imaginário social e cidadania: estudo de caso sobre os ex-ocupantes da via expressa, no município de Florianópolis*. Salvador : sem editora, 2002. Registro no Infohab: nº 92769.

7. NEVES, V. & AZEVEDO, J. *Saneamento básico no município de Teresópolis*. Rio de Janeiro : sem editora, 1997. Registro no Infohab: nº 87716.

8. BENEVIDES, J. R. *A educação ambiental nos programas de saneamento para áreas de baixa renda*. Salvador : sem editora, 1998. Registro no Infohab: nº 88999.

9. AZEVEDO, J.; NASCIMENTO, I. C. A. & MENDES, O. F. *Panorama dos problemas gerados pelos resíduos sólidos urbanos no Brasil.* Rio de Janeiro : sem editora, 2003. Registro no Infohab: nº 88900.

10. KOBAYASHI, M. M. *Processo de cálculo hidráulico de sistemas de drenagem pluvial urbana.* Campo Grande : sem editora, 2003. Registro no Infohab: nº 88942.

11. AZEVEDO, J.; GUIMARÃES, L. T. & MORENO, R. A. N. *Inventário das unidades locais industriais potencialmente poluidoras de Nova Friburgo: um estudo de caso.* Rio de Janeiro : sem editora, 1994. Registro no Infohab: nº 88979.

12. NASCIMENTO, I. C. A. *Construção de um conjunto de indicadores de gerenciamento de resíduos sólidos.* Rio de Janeiro : sem editora, 2003. Registro no Infohab: nº 88992.

13. ULHOA, M. B. *Estudo sobre a implantação de um aterro sanitário intermunicipal no município de Nova Lima, através de uma solução consorciada.* Belo Horizonte : sem editora, 2003. Registro no Infohab: nº 88978.

14. BORANGA, M. L. M. *A influência das variáveis ambientais no valor de unidades habitacionais no município de Campo Grande.* Campo Grande : sem editora, 2003. Registro no Infohab: nº 90281.

15. DANTAS, R. A. *Modelos espaciais aplicados ao mercado habitacional: um estudo de caso para a cidade do Recife.* Recife : sem editora, 2003. Registro no Infohab: nº 91619.

16. ROQUE, J. A. *Sistema construtivo em aço patinável e bloco de concreto celular autoclavado: análise de protótipo de moradia de interesse social.* São Paulo : sem editora, 2003. Registro no Infohab: nº 91734.

17. FERREIRA, E. C. L. *Participação comunitária em xeque: projetos sociais do programa "morar melhor " no Tocantins: a comunidade está participando?* Rio de Janeiro : sem editora, 2002. Registro no Infohab: nº 92343.

18. SALGADO NETO, O. *Indicadores de desempenho na gestão e na regulação dos serviços de saneamento*. São Carlos : sem editora, 2002. Registro no Infohab: nº 92319.

19. TORELLY, L. P. P. *A região metropolitana do Distrito Federal e o Programa Habitar/Brasil-BID*. Brasília : sem editora, 2004. Registro no Infohab: nº 92756.

20. ARAÚJO, N. R. *A importância das transferências do orçamento geral da união para os municípios de Rondônia: dificuldades na operacionalização destes recursos (principais causas)*. Porto Velho : sem editora, 2001. Registro no Infohab: nº 92697.

21. AMARAL, D. F.; TELES, L. G.; TAVARES, M. P. *et al. Resíduos sólidos em Juiz de Fora: o plano diretor de limpeza urbana de 1996 e a gestão atual*. Juiz de Fora : sem editora, 2001. Registro no Infohab: nº 92696.

22. MARQUES, J. A. V. *A utilização de alianças estratégicas no combate do défict habitacional: um estudo de caso do conjunto habitacional "Cidade de Deus" em Sete Lagoas - MG*. Varginha : sem editora, 2002. Registro no Infohab: nº 92682.

23. LIMA, L. F. *Análise do potencial de aplicação da avaliação pós-ocupação no pirograma de arrendamento residencial*. Curitiba : sem editora, 2002. Registro no Infohab: nº 92681.

24. VIEIRA, J. E. G. *Educação ambiental: uma proposta de participação comunitária junto aos programas de desenvolvimento urbano operacional pela*

Caixa Econômica Federal. Goiânia : sem editora, 2001. Registro no Infohab: nº 912698.

25. MIRANDA, C. S. *Proposta de parcelamento e infra-estrutura urbana na bacia do Córrego do Moinho em Cuiabá*. Cuiabá : sem editora, 2002. Registro no Infohab: nº 92695.

26. SILVA, C. M. *A aplicação do modelo de project finance em projetos de infra-estrutura*. Brasília : sem editora, 2002. Registro no Infohab: nº 92683.

27. SIQUEIRA, M. A. M. *Frederick Law Olmsted: "sociedade, natureza e cidade"*. Brasília : sem editora, 2001. Registro no Infohab: nº 92757.

28. SANTOS, E. L. *Inadimplência habitacional em Santa Maria*. Santa Maria : sem editora, 2000. Registro no Infohab: nº 93205.

29. SIQUEIRA, M. A. M. *A participação dos municípios na questão do déficit habitacional*. São Paulo : sem editora. Registro no Infohab: nº 92760.

30. SIQUEIRA, M. A. M. *Público, privado, estatal? há uma nova esfera pública em construção?* Porto Alegre : sem editora. Registro no Infohab: nº 92761.

31. BRITO, R. M. A. *Sistema de gestão ambiental*. São Paulo : sem editora e sem data. Registro no Infohab: nº 92754.

32. VIEIRA, J. E. G. *Programas de desenvolvimento urbano e a participação comunitária pela via de um projeto de educação ambiental*. São Paulo : sem editora e sem data. Registro no Infohab: nº 92781.

33. LIINS, M P. E.; NOVAES, L. F. L.; PAIVA, S. A. *et al. Avaliação imobiliária pelo método da envoltória sob dupla ótica.* São Paulo : sem editora e sem data. Registro no Infohab: nº 92776.

34. MELLO, E.; DUBAIS, B. & LANDIM, R. C. C. *Meio ambiente & ecologia x direito & ética.* São Paulo : sem editora e sem data. Registro no Infohab: nº 92765.

35. AIVERBECK, C. E. *Planta de valores genéricos: necessidade de compromisso com a realidade de mercado.* Florianópolis : sem editora e sem data. Registro no Infohab: nº 92763.

36. PAIVA, S. A. *Problemas de avaliação decorrentes do estatuto da cidade.* São Paulo : sem editora, 2002. Registro no Infohab: nº 92773.

37. SALGADO, V. M. *Lançamento de um empreendimento imobiliário: qual o melhor momento?* São Paulo : sem editora, 2002. Registro no Infohab: nº 92770.

38. TORELLY, L. P. P. *Orientação para análise do roteiro/diagnóstico.* São Paulo : sem editora, 2004. Registro no Infohab: nº 92766.

39. GOTTSCHALG, M. F. S. *Preservação do patrimônio cultural e desenvolvimento urbano.* Belo Horizonte : sem editora, 2000. Registro no Infohab: nº 92767.

40. CARNEIRO, A. M. S. M. *O mercado habitacional europeu.* São Paulo : sem editora, 2001. Registro no Infohab: nº 92780.

41. SIQUEIRA, M. A. M. *Pierre Patte e resíduos sólidos: um breve estudo comparativo do trato de resíduos sólidos ontem e hoje.* São Paulo : sem editora e sem data. Registro no Infohab: nº 92758.

42. SIQUEIRA, M. A. M. *Resíduos sólidos na história: uma breve reflexão sobre o trato dos resíduos sólidos no passado.* São Paulo : sem editora. Registro no Infohab: nº 92759.

43. VASCONCELOS, S. J. *Avaliação ambiental estratégica do Programa Proceder III de Pedro Afonso.* Palmas : sem editora, 2004. Registro no Infohab: nº 94498.

44. QUARESMA NETO, E. *Pesquisas em pequenos sistemas de esgotos sanitários.* São Paulo : sem editora e sem data. Registro no Infohab: nº 92764.

45. FIGUEIREDO, P. M. L. *Protocolo verde: incorporação da variável ambiental no processo de concessão de crédito público.* São Paulo : sem editora e sem data. Registro no Infohab: nº 92768.

46. CAVALCANTE, C. Q. B. *O uso misto na reocupação dos edifícios subutilizados nos centros urbanos.* Recife : sem editora, 2004. Sem número de registro no Infohab.

47. ARAÚJO, J. A. F. *Intervenção urbana em zona de habitação degradada em Ipatimga - MG (assentamento subnormal).* Brasília : sem editora, 2004. Sem número de registro no Infohab.

48. SUGAI, J. J. *O direito constitucional à moradia e os instrumentos jurídicos para sua efetividade.* São Paulo : sem editora, 2003. Sem número de registro no Infohab.

49. AVERBECK, C. E. *Os sistemas de cadastro e planta de valores no município: prejuízos da desatualização.* Florianópolis : sem editora, 2003. Sem número de registro no Infohab.

50. ROMANO, T. P. *Rifiuti & cittadinanza: un' esperienza brasiliana*. Brasília : sem editora, 2004. Sem número de registro no Infohab.

51. RIOS, S. V. *A terceirização dos serviços técnicos de engenharia da Caixa no segmento desenvolvimento urbano*. Brasília : sem editora, 2001. Sem número de registro no Infohab.

52. SOUZA, J. W. R. *A reforma de 1998/1999 da previdência social brasileira*. São Paulo : sem editora, 2000. Sem número de registro no Infohab.

53. SIVIERO, L. A. S. & MENEGÁZ, P. J. M. *Supervisão e controle simplificado das utilidades de um edifício já construído, priorizando a redução de custos operacionais*. Vitória : sem editora, 2000. Sem número de registro no Infohab.

54. FROTA, V. P. N. *Demanda e oferta habitacional na cidade de Manaus na década de 90*. Manaus : sem editora, 1997. Registro no Infohab: nº 92699.

55. TORELLY, L. P. P. *Alternativas e possibilidades para um programa de parcerias em habitação e desenvolvimento urbano*. Brasília : sem editora, 2002. Sem número de registro no Infohab.

56. GALIZA, H. R. S. *A reabilitação urbana dos sítios históricos brasileiros*. Rio de Janeiro : sem editora. 2002. Sem número de registro no Infohab.

ANEXO 3 - Resultados dos testes de precisão realizados com a Base de Dados do Infohab e o Protótipo com aplicação da Mineração de textos

Pesquisa (Formulários)	Resultados com a Base do Infohab (Índice de Precisão)	Resultados com o Protótipo (Índice de Precisão)
Formulário 1 – Luiz Felippe Pinheiro Júnior	0,0%	18,18%
Formulário 2 – Márcia Cambraia Belderain	16,66%	100%
Formulário 3 – Valdi Dantas	50%	36,66%
Formulário 4 – Elizeu Viana Machado Júnior	100%	30,76%
Formulário 5 – Mário Ricardo F. M. Maia	0,0%	0,0%
Formulário 6 – Cláudia Brandão de Serpa	20%	38,88%
Formulário 7 – Ivan Domingues das Neves	0,0%	18,75%
Formulário 8 – Maria Solange Fonseca	50%	70%
Formulário 9 – Emmanuel Braz	50%	75%
Formulário 10 – Márcio de Almeida Machado	0,0%	36,84%
Formulário 11 – Olavo José Perondi	60%	21,42%
Formulário 12 – José Roberto Lopes	15,38%	44,44%
Formulário 13 – Gislaine P. B. de Sá	0,0%	33,33%
Formulário 14 – Raimundo N. de Sousa	50%	21,73%
Formulário 15 – Patricia Marie Jeanne Cormier	33,33%	10,52%
Formulário 16 – Rosane Helena C. P. de Araújo (bibliotecária chefe da CEDIN)	100%	34,61%
Formulário 17 – Elcy Elda Gomes Leão (bibliotecária da CEDIN)	50%	17,85%
Formulário 18 – Márcia Rocha de Aguiar	37,5%	11,11%
Formulário 19 – Cláudia Regina de Paula Mattos (bibliotecária da CEDIN)	42,85%	23,8%
Formulário 20 – Camila Akemi Harada	20%	16,66%
Formulário 21 – Sandra Neves	57,14%	51,16%
Formulário 22 – Júlio César Paixão Lopes	0,0%	72,41%

ANEXO 4 – Carta de apresentação da pesquisa

UNIVERSIDADE DE BRASÍLIA
Faculdade de Economia, Administração, Contabilidade e Ciência da Informação
Departamento de Ciência da Informação e Documentação
70910-900 - BRASÍLIA-DF

Brasília 07 de julho de 2004.

Senhora Coordenadora,

Tenho a satisfação de apresentar-lhe o Professor Rogério Henrique de Araújo Júnior, aluno de doutorado do Programa de Pós-Graduação da Universidade de Brasília, que elegeu como tema de sua tese, o estudo comparado entre a mineração de textos e a indexação manual, por meio da análise do índice de precisão de resposta no processo de busca e recuperação da informação: o caso do Infohab.

Necessitamos de Vosso apoio para a viabilização desta pesquisa dada a importância da utilização do *software* BR/*Search* para o trabalho e a possibilidade de aplicação dos resultados da pesquisa na Centralizadora de Documentação e Informação da Caixa. Desta forma, solicitamos a gentileza de receber o pesquisador supracitado.

Informamos, ainda, que caso seja do interesse desta Coordenação, os resultados do trabalho poderão ser levados ao seu conhecimento durante e após a sua conclusão.

Certos de contar com a valiosa colaboração de Vossa Senhoria, agradecemos antecipadamente.

Profª Drª KIRA TARAPANOFF
Orientadora da Pesquisa

Ilma. Sra.
Coordenadora da Gestão da Informação da Caixa
NIESTA