Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Metodologia para Recomendação de Consultores Ad-Hoc Baseada na Extração de Perfis do Currículo Lattes

Weliton Moreira Bastos

Dissertação apresentada como requisito parcial
para conclusão do Mestrado em Informática

Orientador
Prof. Dr. Marcelo Ladeira

Brasília
2009

CIP — Catalogação Internacional na Publicação

Bastos, Weliton Moreira.

Metodologia para Recomendação de Consultores Ad-Hoc Baseada na Extração de Perfis do Currículo Lattes / Weliton Moreira Bastos. Brasília : UnB, 2009.
114 p. : il. ; 29,5 cm.

Dissertação (Mestrado) — Universidade de Brasília, Brasília, 2009.

1. Recomendação, 2. extração de perfis, 3. filtragem de dados,
4. mineração de dados, 5. mineração de textos

CDU 004.4

Dedicatória

A Jesus Cristo, meu Senhor e Salvador: *"Porque dele e por ele, e para ele, são todas as coisas; glória, pois, a ele eternamente. Amém."* (Romanos 11:36).

Às pessoas mais importantes em minha minha vida: minha amada esposa e minhas duas preciosas filhas filhas.

Agradecimentos

A Deus, fonte da vida, de toda verdade e de todo conhecimento.

A minha esposa que com amor e carinho suportou com paciência minha quase ausência em muitos momentos.

A minhas filhas, dádivas de Deus, pela tolerância com que suportaram a redução de atenção a que foram submetidas.

Ao Dr. Marcelo Ladeira, que me orientou e acompanhou durante toda jornada.

Ao CNPq pelo apoio, sem o qual teria sido impossível a realização deste trabalho.

Sumário

Lista de Figuras

Lista de Tabelas

Resumo

Segundo Han e Caryps (2005), recomendação é uma técnica de filtragem personalizada cujo objetivo é predizer se um usuário vai gostar de um determinado item, ou qual o conjunto de itens são mais relevantes e úteis para um grupo de usuários. A sobrecarga de informações imposta pela Internet e a necessidade de determinar com rapidez e eficiência o que é relevante e útil para os usuários têm feito com que técnicas de recomendação sejam amplamente utilizadas em sistemas baseados na Web.

Técnicas de recomendação estão presentes em muitas situações que como comércio eletrônico, sítios de relacionamento e bibliotecas digitais. A seleção e recrutamento de recursos humanos com base no perfil dos profissionais, é uma área de aplicação que atende às características de sistemas de recomendação, pois consiste em identificar quais os profissionais cujos perfis são mais adequados à execução de um conjunto de tarefas.

Um caso particular de seleção de recursos humanos é a indicação de consultores para avaliação de projetos. Nesse caso, deve-se identificar quais os profissionais com qualificações mais adequadas para avaliação dos projetos com base na similaridade entre os perfis dos consultores e dos projetos.

Sistemas de recomendação de consultores devem levar em conta os perfis do consultores, do proponentes e do projetos a serem avaliados, além de possuir mecanismos para detectar e minimizar possíveis conflitos de interesses que tornariam as avaliações suspeitas.

Este trabalho propõe uma metodologia para recomendação de consultores para avaliação de projetos no âmbito do Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq, aplicando estratégias de filtragem baseada em conteúdo. Esta metodologia utiliza o modelo de espaço vetorial (VSM - *vector space model*) para determinar o grau de semelhança entre os perfis dos consultores e proponentes e entre os perfis dos consultores e projetos.

Palavras-chave: Recomendação, extração de perfis, filtragem de dados, mineração de dados, mineração de textos

Abstract

According to Han and Caryps (2005), recommendation is a customized filtering technique whose goal is predict whether a user will like a particular item, or what set of items are most relevant and useful to a group of users. The overload of information imposed by the Internet and the need to determine quickly and efficiently what is relevant and useful to the users have done with that recommendation techniques are been widely used in systems based on Web.

Recommendation techniques are presents in many situations such as electronic commerce, social networking websites and digital libraries. The selection and recruitment of human resources based on the profiles of professionals, is one application area that meets the requirements of recommendation systems, since it consists in identifying the professionals whose profiles are most suitable for the implementation of a set of tasks.

A particular case of selection of human resources is an indication of consultants for evaluation of projects. In this case, must identify practitioners with skills more appropriate for evaluating projects based on the similarity between the profiles of consultants and projects.

Recommendation systems of consultants should consider the consultants' profiles, the proponents' profiles and projects' profiles to be evaluated, and have mechanisms to detect and minimize possible conflicts of interest that would make the evaluations suspicions.

This paper proposes a methodology for the recommendation of consultants for

project evaluation under the National Council for Scientific and Technological Development - CNPq, applying strategies based filtering content. This methodology uses the vector space model (VSM - vector space model) to determine the degree of similarity between the profiles of consultants and bidders and between the profiles of consultants and projects.

Keywords: Recommendation, role extraction, data filtering, data mining, text minning

Capítulo 1

Introdução

Este capítulo apresenta a definição do problema abordado, os objetivos gerais e específicos do projeto, as áreas do conhecimento envolvidas e as contribuições esperadas ao final do trabalho.

Han e Carypis [Han and Karypis, 2005] definem sistemas de recomendação como uma *"tecnologia de filtragem de informação personalizada usada para predizer quando um usuário específico vai gostar de um item em particular (problema da predição) ou para identificar um conjunto de N itens que serão de interesse de certos usuários (problema das N melhores escolhas)"*. Em outras palavras, recomendação consiste em fornecer a terceiros informações, produtos ou serviços que sejam relevantes para quem as recebe, no contexto no qual são realizadas. Uma recomendação pode ser solicitada pelo usuário, ou pode simplesmente ser oferecida sob a hipótese de que a pessoa a quem se destina a sugestão necessita, deseja ou vai se interessar pelo que está sendo oferecido.

Encontramos esse tipo de comportamento em nossos relacionamentos interpessoais quando, por exemplo, sugerimos a alguém que compre algo, leia um livro ou que assista a um filme. O mesmo pode ser observado em sistemas de comércio eletrônico, serviços de bibliotecas, sítios de relacionamentos, ferramentas de busca na rede mundial de computadores, bem como nas indicações de filmes e espetáculos

realizadas por especialistas através dos meios de comunicação.

A recomendação pode ser realizada por um ser humano, como um crítico de cinema, um enólogo, um parente ou um amigo. Nesses casos, a experiência pessoal, o conhecimento prévio e o relacionamento entre as partes envolvidas são fatores subjetivos que influenciam na forma como a recomendação é realizada e em como é percebida pela outra parte. A credibilidade de quem faz a recomendação e outros aspectos psicológicos ainda mais complexos vão afetar a maneira como essa recomendação será recebida e acatada ou rejeitada.

Os sistemas de recomendação automática tentam aproximar o comportamento da máquina dessa habilidade humana. Para isso utilizam metodologias de filtragens que caracterizam o comportamento do sistema conforme o foco seja colaborativo, baseado no conteúdo, baseado em regras ou híbrido - nesse caso, uma mistura de colaborativo e baseado em conteúdo.

Na filtragem colaborativa, os próprios usuários fornecem as informações que são necessárias para o funcionamento do sistema de forma explícita, ou implícita. Na modalidade explícita, isso é feito pelo preenchimento de questionários de avaliação e preferências, ou por meio de indicações na qual um usuário recomenda diretamente um produto ou serviço para outro usuário.

A filtragem colaborativa explícita depende da disposição do usuário em responder perguntas, inscrever-se em grupos de interesse, fóruns e comunidades, ou em realizar indicações diretamente no sistema para um amigo ou colega. Essa última modalidade é especialmente influenciada pela credibilidade da pessoa que realiza a recomendação, principalmente nos meios acadêmicos, científicos e profissionais.

A filtragem colaborativa implícita é resultado de se manter um registro histórico das ações dos usuários, e de se aplicar sobre essa base de informações técnicas de mineração de dados e mineração de textos. Isso permite identificar tendências, padrões de comportamento e grupos de interesses dos usuários e redes sociais que esses usuários participem explícita ou implicitamente. Essa abordagem de

filtragem têm a vantagem de não requerer nenhuma ação específica por parte do usuário, a não ser utilização do sistema. As técnicas de filtragem colaborativa permitem a construção de recomendações do tipo *top-N* (os N mais lidos, acessados, ouvidos, recomendados, ...) e *cross-sell* (quem se interessou por X também se interessou por Y) [Shahabi and Chen, 2003].

A filtragem baseada em conteúdo procura identificar qual item é mais adequado aos usuários que possuem um determinado perfil, baseado nas características dos itens a serem recomendados. Nesse caso, é possível aplicar técnicas de mineração de dados para identificar grupos de interesses e classes de usuários.

A recomendação automática também pode ser baseada em regras. Por exemplo, em um site de comércio eletrônico, se um usuário adquirir uma máquina fotográfica digital, o sistema pode oferecer um estojo para transporte da máquina, um cartão de memória adicional ou uma impressora especial para fotografias. A dificuldade dessa abordagem é que todas as regras devem estar programadas no sistema, ou devem ser configuráveis. O procedimento de alteração das regras é oneroso, requer um conhecimento especializado e é pouco flexível; não é capaz de aprender, descobrir tendências nem tirar vantagens do comportamento de grupo exibido pelos usuários. A recomendação baseada em regras pode ser combinada com as outras abordagens já mencionadas.

Este trabalho concentra-se na recomendação automática para seleção de recursos humanos, particularmente na recomendação de consultores para avaliação de projetos. O objetivo principal é identificar e sugerir pessoas que possuam experiências, habilidades e talentos específicos para exercer uma determinada função ou realizar uma tarefa específica. Essa seleção pode ser realizada visando a contratação de um profissional para ocupar um cargo ou função, liderar um projeto, prestar consultoria, ou escolher um funcionário para realizar uma tarefa pré-definida e assim por diante. Para tanto, o sistema deve manter um banco de dados contendo currículos atualizados dos potenciais candidatos a recomendação, descrição dos requisitos

que os candidatos devem atender, características das funções e tarefas executas na empresa, histórico das contratações anteriores, resultados anteriores, diretrizes políticas da empresa contratante e assim por diante.

Este trabalho foca especificamente a seleção de consultores avaliadores de propostas de projetos no contexto do CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico. A seleção de consultores avaliadores pertence ao escopo seleção de recursos humanos. No caso particular do CNPq, a indicação de consultores é parte do processo de julgamento de propostas de projetos. Uma proposta de projeto visa a obtenção de recursos de fomento para financiamento de: projetos de pesquisa, bolsas de estudo, bolsas de pesquisa, apoio a realização de eventos, apoio a editoração e auxílio viagem para participação em eventos [CNPq, 2007].

A recomendação consultores deve basear-se nos perfis dos consultores disponíveis para recomendação, nos perfis dos proponentes dos projetos, na ação dos consultores em aceitar ou rejeitar a indicação, na ação dos técnicos do CNPq ao indicar consultores previamente recomendados, nas características do projeto e em informações que dependem do contexto específico no qual as recomendações são realizadas (regras do sistema).

Os critérios de similaridade para recomendar um consultor podem ser positivos, negativos ou excludentes. Critérios positivos são aqueles que mantêm uma relação direta com a probabilidade de a recomendação ser realizada, ao passo que os negativos são aqueles que mantêm uma relação inversa. Os critérios excludentes são impeditivos para a recomendação independentemente do grau de similaridade indicados pelos demais critérios.

A diferenciação entre critérios de similaridade positivos, negativos e excludentes é necessária para reduzir a probabilidade de recomendação de consultores que possuam conflitos de interesses em relação ao objeto de avaliação. Por exemplo, um consultor que tenha submetido projeto concorrente com o projeto que ele mesmo vai avaliar, torna-o interessado nos resultados e, portanto, suspeito para emitir pare-

cer, logo ele não deve ser recomendado. Por outro lado, um consultor pode ter maior ou menor grau de proximidade com os proponentes. Isso pode variar desde vínculos diretos como os membros da equipe de projeto; o consultor pode possuir produção científica ou tecnológica conjunta com o proponente ou o consultor e o proponente podem ter um relacionamento orientador-orientando. Essa lista de relacionamentos pode evoluir para situações mais vagas como possuir vínculo com a mesma instituição no mesmo departamento e na mesma cidade. Nesses casos, a proximidade pode não impedir a recomendação mas apenas reduzir a probabilidade de sua recomendação automática.

Dentre as áreas de pesquisa relacionadas com este trabalho, destacam-se a Ciência da Computação e Ciência da Informação, mais especificamente: banco de dados (armazenamento e recuperação de informação - RI), inteligência artificial (representação do conhecimento, aprendizagem automática, processamento de linguagem natural, mineração de textos), algoritmos de busca, técnicas de recomendação, gestão da informação e do conhecimento e arquitetura da informação.

No campo da Psicologia e da Administração, tem-se a área de recrutamento e seleção de pessoal, que se baseia na análise de currículos, entrevistas e aplicação de testes. Nesse contexto, a seleção automática de candidatos com maior probabilidade de atender os requisitos demandados pode reduzir o número de entrevistas e testes que seriam aplicados desnecessariamente.

1.1 Objetivo Geral

Propor uma metodologia de recomendação consultores *ad-hoc* para avaliar propostas de projetos de pesquisa submetidos ao CNPq, baseada na extração de perfis dos Currículos Lattes dos proponentes e dos consultores e nos perfis das propostas de projetos.

1.1.1 Objetivos Específicos

Avaliar diferentes formas de uso dados dos currículos para composição dos perfis dos pesquisadores e das propostas:

- palavras-chave dos currículos dos pesquisadores e palavras-chave dos projetos,
- termos da produção científica e tecnológica dos currículos dos pesquisadores e termos extraídos dos projetos, e
- termos da última formação dos pesquisadores e termos extraídos dos projetos.

1.2 Áreas de Pesquisas Relacionadas

Neste trabalho vamos abordar explicitamente as áreas de recomendação e filtragem híbrida conteúdo-colaborativa e técnicas de mineração de textos.

1.3 Contribuição

Espera-se, através deste trabalho, contribuir para o desenvolvimento das técnicas de extração de perfis e para o seu uso em sistemas de recomendação.

Do ponto de vista tecnológico será delinear um modelo de uma aplicação para extração de perfis e recomendação de consultores *ad-hoc* para uso no ambiente de produção no CNPq e que possa ser adaptado a outros contextos.

1.4 Organização deste Documento

Este documento está organizado da seguinte forma: o capítulo dois apresenta a fundamentação teórica relacionada com recomendação e recuperação de informação, com ênfase no modelo de espaço vetorial e suas variantes mais importantes.

O capítulo três apresenta alguns exemplos de sistemas de recomendação propostos na bibliografia consultada: Yoda, Implicit, W-REMAS, Sistema de Recomendação para Bibliotecas Digitais e uma proposta de uso de recomendação para recuperação de perfis de usuários do Currículo Lattes.

O capítulo quatro apresenta o sistema de recomendação de consultores em uso no CNPq, suas principais características, vantagens, desvantagens e uma análise de desempenho do mesmo.

O capítulo cinco detalha a metodologia proposta, apresenta os pressupostos da metodologia e os critérios de similaridade.

O capítulo seis contém os resultados dos experimentos realizados, utilizando três conjuntos de dados textuais para construção da representação dos perfis e apresenta uma analise do desempenho das abordagens propostas em comparação com o sistema atual e as principais dificuldades encontradas.

O capítulo sete apresenta as conclusões e sugestões para desenvolvimentos futuros e para superação das dificuldades encontradas.

Capítulo 2

Fundamentação teórica

Este capítulo discute as principais abordagens de recomendação automática, suas vantagens e desvantagens. Apresenta conceitos relacionados com os principais formalismos utilizados na proposta para seleção de perfis de consultores: indexação automática baseada no modelo de espaço vetorial - VSM (do inglês *vector space model*) e suas principais variações.

2.1 Recomendação

O uso de recomendação faz parte do dia-a-dia de todas as pessoas, principalmente diante de situações novas como a compra de um novo modelo de equipamento eletrônico, escolha de um filme, ou elaboração de um roteiro de férias. A lista de possibilidades é extensa, mas em todas as situações o comportamento das pessoas é semelhante: o primeiro passo é a pesquisa de informações que possam embasar a decisão. Essas informações incluem a opinião de outras pessoas, sejam elas especialistas ou não. Nesse caso, é o interessado quem solicita a informação, em outras situações acontece o contrário: a informação é oferecida sob alegação de que será útil sem que haja solicitação por parte daquele a quem se destina. Por exemplo, ao ler um livro, assistir a um filme ou comprar um produto, frequentemente a pessoa

se lembra de alguém que ”certamente vai gostar” daquilo. Como resultado, uma recomendação direta não solicitada é endereçada ao suposto interessado.

Em todos os casos o que está sendo demandado, ou oferecido, é informação que supostamente deverá ajudar, ou induzir, alguém a tomar uma decisão por este ou aquele produto, serviço, atividade, etc. Informação que se supõe relevante, útil e até mesmo necessária àquele a quem se destina.

Com o crescimento do volume de informação disponível na Internet e o desenvolvimento do comércio eletrônico, a utilização de mecanismos de recomendação torna-se cada dia mais relevante. Uma simples consulta em qualquer mecanismo de busca, pode retornar milhares de resultados, até mesmo milhões.

É virtualmente impossível para qualquer pessoa visualizar sempre todos os resultados de uma consulta em busca do que é do seu interesse em meio ao que pode ser apenas lixo. A solução é filtrar informações de tal forma que o usuário receba primeiramente aquelas que são mais relevantes no seu próprio contexto. Isso permitiria uma redução de tempo e esforço realizado pelo usuário na tentativa de encontrar o que procura e, no caso do comércio eletrônico, aumentaria as vendas ao apresentar ao possível comprador itens que provavelmente são do interesse dele. Devido a essas características, os sistemas de recomendação automáticos estão crescendo em importância. Isso pode ser observado com facilidade em qualquer sítio de busca, de comércio eletrônico ou de relacionamentos.

2.1.1 Recomendação automática

Sistemas de recomendação automática são relativamente novos e apresentam desafios ainda não resolvidos, tais como o problema básico de aprendizagem, que consiste em predizer as ações ou o interesse de um grupo de usuários a partir da observação de seu comportamento [Birukov et al., 2005] e a determinação das N melhores escolhas (*top-N*) que sejam relevantes para um usuário em um contexto específico [Han and Karypis, 2005].

O crescimento vertiginoso do volume e da variedade de dados nos atuais sistemas de informação, bem como a "sobrecarga de informação" que é imposta pela Internet, fazem com que a utilização de estratégias de recomendação sejam de grande relevância em contextos como comércio eletrônico, sítios de relacionamentos, bibliotecas digitais, motores de busca e muitos outros. O uso de técnicas de recomendação permite que resultados melhores sejam identificados mais rapidamente, evitando que o usuário tenha que navegar através de centenas, ou milhares, de páginas recebidas em resposta a uma consulta.

Pesquisando no Google por "inteligência artificial", por exemplo, o resultado obtido foi "aproximadamente 208.000.000" resultados, dos quais ele personalizou dez que julgou ser do interesse do usuário e ainda exibiu um conjunto de "pesquisas relacionadas" ao argumento de busca submetido, como pode ser visto na figura 2.1.

Este exemplo ilustra o uso de estratégias de recomendação. Não se trata de realizar uma busca segundo algum critério e apresentar os resultados para o usuário, mas de tentar inferir o que é mais adequado para aquele usuário naquele momento e apresentar esses resultados a ele, levando em conta o perfil do usuário, suas preferências explícitas e implícitas, seu comportamento, as preferências dos grupos de interesse e comunidades de afinidades das quais o usuário pode ser considerado membro, além de outros critérios que dependem da aplicação, do produto ou serviço a ser oferecido e do contexto específico em que as transações ocorrem.

A realização de uma recomendação deve levar em conta a relevância daquilo que está sendo recomendado do ponto de vista do usuário. Isso por si só é um problema extremamente complexo e que permanece em aberto, pois o que é relevante para alguém em um contexto não será necessariamente relevante para outra pessoa no mesmo contexto. Por outro lado, o que é relevante para uma pessoa em uma determinada situação pode não ser relevante para essa mesma pessoa em situação semelhante em outro momento.

Figura 2.1: Página de consulta ao Google

Segundo [Porter, 2006], a filtragem de informação baseada em técnicas de recomendação possui as seguintes vantagens:

- é baseada na atividade real dos usuários;
- possibilita a descoberta de novas relações não declaradas;
- permite personalização dos resultados;
- o sistema está sempre atualizado;
- redução de esforço organizacional para manter ontologias e taxonomias, pois a recomendação automática baseia-se em fatos acumulados na relação do usuário com a empresa.

Marques (2007) acrescenta a essa lista que, quando o universo a ser consultado é desconhecido, ou grande ao ponto de tornar proibitiva a navegação através de todos os registros recuperados, o uso de recomendação tem vantagens evidentes ao recuperar os primeiros registros que provavelmente são mais relevantes para o usuário.

Potter (2006) lista também as desvantagens do uso de sistemas de recomendação automática:

- dificuldade para manter atualizados os dados históricos por causa do grande volume de registros;
- manutenção do sistema de recomendação;
- possibilidade de recomendações falhas devido aos relacionamentos não declarados pelos usuários, mas de alguma forma mapeados pelo mecanismo de determinação de similaridades;
- usuários que brincam com o sistema, provocando distorções nas recomendações.

Sistemas de recomendação devem considerar três tipos de informação: os itens a serem recomendados, os usuários aos quais as recomendações se destinam e informações transacionais sobre o comportamento dos usuários ao longo de um determinado período de tempo.

Os sistemas de recomendação também podem ser baseados em conhecimento. Nesse caso, um especialista, ou administrador do sistema, define regras para recomendação. Essas regras podem ser baseadas em conhecimento acumulado pelo especialista ou administrador, podem ser obtidas por técnicas de mineração de dados, ou de textos, ou podem ser frutos de políticas da empresa para aumentar as vendas, ou para aumentar o acesso a informações sobre determinados produtos.

As abordagens dos sistemas de recomendação dependem de como essas informações são utilizadas: a filtragem baseada no conteúdo é focada nos itens a serem recomendados combinados com os perfis dos usuários. A filtragem colaborativa é baseada na iteração do usuário com o sistema, podendo dispor de avaliações explícitas dos itens e do histórico de iteração dos usuários. A terceira abordagem é uma combinação das duas primeiras. As estratégias baseadas no conteúdo realizam filtragens ou classificação dos itens com base em características que de alguma forma se relacionam com o perfil dos usuários.

Na filtragem colaborativa explícita o usuário é solicitado a avaliar os produtos, ou perfis de outros usuários. Marques (2007) sugere que as opiniões dos usuários não podem ser consideradas uniformemente iguais em qualquer contexto, pois há situações nas quais a reputação do usuário ou sua qualificação deve ser considerada no processo de recomendação. Por exemplo, na recomendação de currículos de pesquisadores, a opinião dos pesquisadores mais renomados deve ser considerada mais importante, ou na avaliação de artigos científicos, aqueles pesquisadores que possuem produção no domínio do conhecimento envolvido devem ter uma opinião mais relevante do que aqueles que não têm.

A filtragem baseada em conhecimento é relativamente simples de ser implementada, mas não de ser mantida, pois requer atualização constante da base conhecimento e é difícil de automatizar, principalmente se a fonte do conhecimento derivar da experiência de especialistas responsáveis pelas regras, ou se as regras forem oriundas de políticas da empresa. Generalizar sistemas que utilizam essa abordagem é bastante complicado, uma vez que as regras e forma como são utilizadas dependem do contexto nos quais são utilizados. A inclusão ou exclusão de novas regras demandam interferência humana, tanto na concepção quanto na implementação.

Chen e Shahabi (2003) afirmam que a filtragem baseada em conteúdo é criticada por sua limitação de conteúdo, geralmente restrita a determinados tipos ou aspectos extraídos dos itens. Além disso, padece de super-especialização, isto é, baseia-se unicamente no conteúdo dos perfis dos usuários e não permite que sejam explorados novos itens que não estejam relacionados com esses perfis. Afirmam ainda que a filtragem colaborativa resolve esses problemas, entretanto introduz outros problemas:

escalabilidade – o tempo necessário para determinar os conjuntos de similaridades cresce linearmente com o número de itens e de usuários;

dados esparsos – os usuários relutam em fornecer informações, produzindo uma distribuição esparsa de características nos perfis, levando o sistema a realizar recomendações imprecisas;

sinonímia – desconsidera associações latentes entre os itens por ignorar suas características, como resultado muitos deles não são recomendados, introduzindo falsos negativos.

Para resolver esses problemas, diversas técnicas têm sido propostas como redução dimensional, divisão em classes e redes bayesianas. Essas técnicas reduzem o problema da escalabilidade ao extraírem padrões por meio de um processamento em lote para uso em tempo real, entretanto reduzem a acurácia e aumentam a complexidade das realização das recomendações em tempo real proporcionalmente ao número de classes envolvidas. Para redução dos problemas de sinonímia e de dados esparsos, técnicas baseadas em regras de associação e categorização são aplicadas aos registros históricos com objetivo de captar associações latentes que são combinadas com as colaborações dos usuários para produzir novas recomendações. Isso faz com que o tempo de processamento cresça proporcionalmente ao volume de dados agregados [Shahabi and Chen, 2003].

Apesar das dificuldades e limitações, mais e mais sistemas estão incorporando recomendações automáticas ao seu repertório comportamental, principalmente em sistemas de comércio eletrônico, onde oportunidades de venda precisam ser criadas no momento exato em que o usuário esteja propício.

A tabela 2.1 resume algumas dessas abordagens, indicando suas principais vantagens e desvantagens.

2.2 Modelo de espaço vetorial

Segundo Salton apud Polyvanyy e Kuropa (2007), o modelo de espaço vetorial foi usado para indexação e busca de documentos pela primeira vez no sistema de re-

Abordagens de recomendação	**Vantagens**	**Desvantagens**
Especialista humano	Flexível. Preciso. Simples.	Não é automatizável. Requer muitos especialistas. Tempo para registrar as recomendações é elevado.
Baseada em regras	Automatizável. Simples. Eficiente. Consumo baixo de memória.	Dificuldade para incluir novas regras. Dificuldade para generalizar.
Baseada em conteúdo	Permite aplicação de *data mining* para detecção de tendências. Permite identificar comportamentos de grupos. Objetos novos podem ser recomendados. Flexível. Automatizável.	Requer grandes volumes de informação armazenada. Depende de cadastro prévio detalhando dos objetos recomendáveis. Depende de cadastro dos perfis dos usuários.
Colaborativa explícita	Permite identificar comportamentos de grupo. Permite aplicação de *data mining* para detecção de tendências. Flexível. Automatizável.	Requer armazenamento de grandes volumes de informação. Depende de cadastro dos perfis dos usuários. Objetos novos não serão recomendados. Pode ter resultados falseados pelos usuários. Depende de o usuário preencher formulários e responder perguntas.
Colaborativa implícita	Baseada no comportamento real do usuário e não em suas afirmações. Não depende de o usuário preencher formulários ou responder perguntas. Permite identificar comportamentos de grupo. Automatizável.	Requer armazenamento de grandes volumes de informação. Depende de cadastro dos perfis dos usuários.
Social	Permite aplicação de *data mining* para detecção de tendências. Baseado no comportamento real do usuário e não em suas afirmações. Não depende de o usuário preencher formulários e responder perguntas. Automatizável. Flexível.	Requer armazenamento de grandes volumes de informação. Depende de cadastro dos perfis dos usuários.
Híbrida	Depende de como as características de cada abordagem são empregadas. Automatizável. Flexível.	Depende de como as características de cada abordagem são empregadas. Difícil implementação.

Tabela 2.1: Abordagens de recomendação

cuperação de informação SMART desenvolvido pela Cornell University em 1960. Esse modelo baseia-se em uma estrutura algébrica denominada espaço vetorial.

Recio-Garcia e colaboradores (2008) consideram que o modelo de espaço vetorial é uma ferramenta de recuperação de informação de fundamentação estatística com pouco poder de expressão semântica e que apresenta dificuldades para explicar os resultados recuperados, mas concordam que essa técnica apresenta bons resultados, principalmente se combinada com outras técnicas, como por exemplo o modelo booleano de recuperação de informação, agrupamento dos documentos em tópicos de acordo com o assunto de cada um, *LSI - Latent Semantic Index* (*LSI*), ou *Latent Semantic Analisys* (*LSA*) e *Singular Value Decomposition* (*SVD*). Para mais detalhes consulte [Manning et al., 2008] e [Mendes et al., 2002].

Um espaço vetorial $\mathcal{V}$ sobre um corpo $\mathcal{C}$, é um conjunto não vazio de vetores $\mathcal{V}$ e um conjunto de escalares de $\mathcal{C}$ dotados de uma operação de adição de vetores, adição de escalares, multiplicação de escalares e multiplicação de vetor por escalar. Além disso, a adição de vetores é associativa, comutativa, possui elemento neutro e oposto para todo vetor. A multiplicação por escalar é associativa e distributiva em relação a adição de vetores e possui elemento neutro. A multiplicação por escalar é distributiva em relação a adição de escalares [Gonçalves and Souza, 1977].

Um corpo é um conjunto com pelo menos dois elementos distintos (zero e um) dotado das operações de adição e multiplicação, tais que a adição é associativa, comutativa, possui elemento neutro (zero) e todo elemento do corpo possui oposto. A multiplicação é distributiva em relação à adição, é associativa, comutativa, possui elemento neutro (um) e todo elemento diferente de zero possui inverso multiplicativo [Monteiro, 1974].

No escopo de recuperação de informação, é de interesse particular espaços vetoriais sobre o números reais $\Re$. Um espaço vetorial n-dimensional $\Re^n$ é composto por n-uplas na forma $\vec{v} = (c_1, c_2, \ldots, c_n)$, onde $c_i \in \Re$, $i \in \{1, 2, \ldots, n\}$.

O produto interno, ou produto escalar, de dois vetores $\vec{v}_1 = (a_1, a_2, \ldots, a_n)$ e $\vec{v}_2 = (b_1, b_2, \ldots, b_n)$ é definido por:

$$\vec{v_1} \cdot \vec{v_2} = a_1 b_1 + a_2 b_2 + \ldots + a_n b_n \tag{2.1}$$

A norma ou comprimento de um vetor $\vec{v} = (c_1, c_2, \ldots, c_n)$ é dada por

$$|\vec{v}| = \sqrt{c_1^2 + c_2^2 + \ldots + c_n^2} \tag{2.2}$$

Demonstra-se que a relação do ângulo entre dois vetores com o produto escalar é dada por

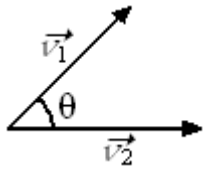

Figura 2.2: Ângulo entre dois vetores

$$\vec{v_1} \cdot \vec{v_2} = |\vec{v_1}| \cdot |\vec{v_2}| \cdot \cos\theta \tag{2.3}$$

onde θ é o ângulo entre os vetores $\vec{v_1}$ e $\vec{v_2}$, assim

$$\cos\theta = \frac{\vec{v_1} \cdot \vec{v_2}}{|\vec{v_1}| \cdot |\vec{v_2}|} \tag{2.4}$$

Para $0 \leq \theta \leq \Pi \Rightarrow 1 \geq \cos\theta \geq -1$, de forma que quanto menor o ângulo entre os vetores envolvidos, maior o cosseno do ângulo entre eles. Pode-se tomar o cosseno do ângulo como uma medida de proximidade entre os vetores, de forma que quanto maior o cosseno do ângulo entre os vetores, menor o ângulo entre eles. Se o produto escalar de dois vetores for igual a zero, os vetores são ditos ortogonais.

Um conjunto de vetores $\mathcal{W} = \{\vec{v_1}, \vec{v_2}, \ldots, \vec{v_k}\}$ é dito linearmente independente, ou simplesmente independentes se, e somente se, a única solução possível para a

equação vetorial $a_1\vec{v_1} + a_2\vec{v_2} + \ldots + a_k\vec{v_k} = \vec{0}$, onde $\vec{0} = (0, 0, ..., 0)$, é a solução trivial $a_1 = a_2 = \ldots = a_k = 0$. Em outras palavras: um conjunto não vazio de vetores $\mathcal{W}$ é linearmente independente, se e somente se, nenhum vetor de $\mathcal{W}$ pode ser escrito como combinação linear dos demais vetores.

Todo espaço vetorial $\mathcal{V}$ pode ser representado por um subconjunto mínimo de vetores de $\mathcal{V}$, digamos $\mathcal{W}$ = $\{\vec{v_1}, \vec{v_2}, \ldots, \vec{v_k}\}$, convenientemente escolhidos tal que qualquer vetor de $\mathcal{V}$ pode ser representado por uma combinação linear única dos vetores de $\mathcal{W}$. Um conjunto $\mathcal{W}$ com essas características é denominado uma base para $\mathcal{V}$, além disso, pode-se provar que $\mathcal{W}$ é linearmente independente. O número de vetores de $\mathcal{W}$ é uma base do espaço vetorial $\mathcal{V}$. Prova-se que todas as bases de $\mathcal{V}$ tem o mesmo número de vetores, esse número é denominado dimensão do espaço vetorial $\mathcal{V}$. Para um vetor qualquer $\vec{v} \in \mathcal{V}$, existem coeficientes reais $a_1, a_2, ...a_n$, tais que $\vec{v} = a_1\vec{v_1} + a_2\vec{v_2} + \cdots + a_n\vec{v_n}$. Nessas condições a n-upla $(a_1, a_2, ...a_n)$ e chamada coordenadas de $\vec{v}$ na base $\mathcal{W}$.

Um vetor é dito normal se seu comprimento for igual a 1. Para qualquer vetor não nulo $\vec{d}$ seu equivalente normalizado é dado por $\vec{\delta} = \dfrac{\vec{d}}{|\vec{d}|}$, tem a mesma direção e sentido que $\vec{d}$. Além disso, todos os vetores de mesma direção sentido possuem a mesma representação normalizada.

O Modelo de Espaço Vetorial - VSM (*Vector Space Model*), pressupõe que é possível extrair um conjunto de termos dos documentos que serão indexados, e que esse conjunto de de termos pode ser usado para construir um espaço vetorial onde cada documento do conjunto pode ser representado por um vetor em um espaço n-dimensional de termos. Dessa forma, a representação vetorial de um documento seria sua coordenada nesse espaço.

Se d é um documento, sua representação vetorial $\vec{d}$ é uma n-upla de números reais $\vec{d}$ = $(t_{d,1}, t_{d,2}, \ldots, t_{d,n})$, onde cada número real $t_{i,d}$ indica a pertinência do termo t_i para representar d. Se $t_{i,d} = 0$, então o termo t_i é irrelevante na representação de d no modelo. O uso dos valores discretos 0 e 1 para os t_i permitem representar

ausência (0) e presença (1) do termo no documento e possibilita a realização de consultas booleanas sobre o modelo. O uso de valores reais dentro de um intervalo permite indicar o grau de pertinência do termo t_i para representar d por sua vez permite consultas mais sofisticadas.

Na sua forma original o modelo VSM é denominado W-VSM (*Word Vector Space Model*) e armazena uma representação dos termos tais como estão no texto sem nenhuma alteração [Ikehara et al., 2001]. Para redução da dimensão da base de vetores e por não contribuírem com as operações de busca e classificação, as termos com frequência elevada e baixa expressividade não são consideradas na construção do VSM, por exemplo: artigos, preposições, numerais, etc.

A Figura 2.3 ilustra três 'documentos': "carro rápido", "carro vermelho" e "carro vermelho rápido". O ângulo θ indica a similaridade entre "carro rápido" e "carro vermelho rápido". A base do espaço vetorial usada para representar os documentos do gráfico é composta por três vetores:

"carro" = (1, 0 , 0)

"rápido" = (0, 1, 0)

"vermelho" = (0, 0, 1)

Qualquer documento nesse espaço será representado por uma combinação linear dos elementos da base, por exemplo:

"carro rápido" = **1** . (1, 0, 0) + **1** . (0, 1, 0) + **0** . (0, 0, 1).

As coordenadas do vetor que representa esse documento ("carro rápido") na base do exemplo acima é dada pelos coeficientes em destaque, que aparecem multiplicando os vetores da base na combinação linear, nesta ordem.

"carro rápido" = (1, 1, 0)

Da mesma forma pode-se escrever as coordenadas dos outros dois documentos do *corpus* do exemplo:

“carro vermelho rápido” = **1**.(1, 0, 0) + **1**.(0, 1, 0) + **1**.(0, 0, 1) = (1, 1, 1)

“carro vermelho” = **1**.(1, 0, 0) + **0**.(0, 1, 0) + **1**.(0, 0, 1) = (1, 0, 1)

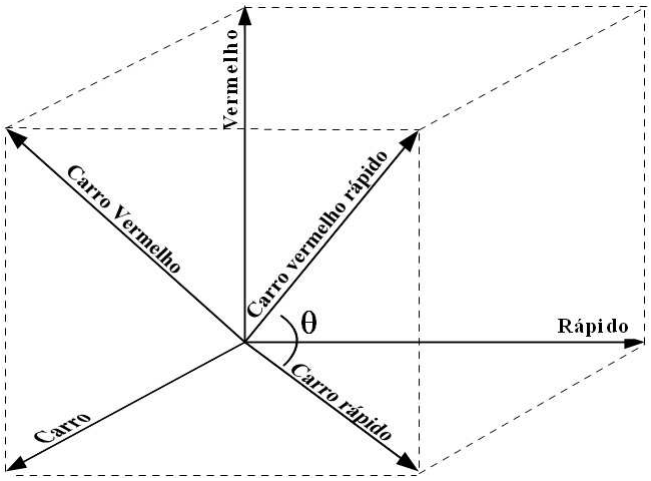

Figura 2.3: Vetores de termos

Fonte: [Polyvyanyy and Kuropka, 2007] p. 7

Durante uma consulta, o argumento de busca é convertido em um vetor-consulta. Esse vetor é expresso na mesma base utilizada para representar os documentos. O vetor-consulta é comparado com os vetores que representam dos documentos armazenados. O conjunto dos vetores mais “próximos” (maiores cossenos) do vetor-busca consiste na resposta à consulta.

Por exemplo, para realizar uma busca pelo documento “veículo rápido e vermelho”, serão necessários os seguintes passos:

1. descarte dos termos não representativos: “e”;
2. extração dos termos a serem pesquisados: “veículo”, “rápido”, “vermelho”;
3. descarte dos termos que não constam na base: “veículo”;
4. obtenção do vetor-consulta através da representação do documento (d_x) a ser procurado na base do VSM:

$$\vec{d_x} = \mathbf{0}.(1,0,0) + \mathbf{1}.(0,1,0) + \mathbf{1}.(0,0,1) = (0,1,1);$$

5. cálculo da similaridade (**Sim**) do vetor-consulta com os vetores que representam os documentos do *corpus*:

$$d_1 = \text{“carro rápido”}$$

$$d_2 = \text{“carro vermelho”}$$

$$d_3 = \text{“carro vermelho rápido”}$$

$$d_x = \text{“veículo rápido vermelho”}$$

$$Sim(d_1, d_x) = \frac{(1,1,0).(0,1,1)}{|(1,1,0)|.|(0,1,1)|} = \frac{1}{\sqrt{2}\sqrt{2}} = \frac{1}{2}$$

$$Sim(d_2, d_x) = \frac{(1,0,1).(0,1,1)}{|(1,0,1)|.|(0,1,1)|} = \frac{1}{\sqrt{2}\sqrt{2}} = \frac{1}{2}$$

$$Sim(d_3, d_x) = \frac{(1,1,1).(0,1,1)}{|(1,1,1)|.|(0,1,1)|} = \frac{2}{\sqrt{3}\sqrt{2}} = \frac{2}{\sqrt{6}} \cong 0,82$$

Considerando somente o resultado mais semelhante ao argumento de busca, a consulta resultaria em “carro vermelho rápido”.

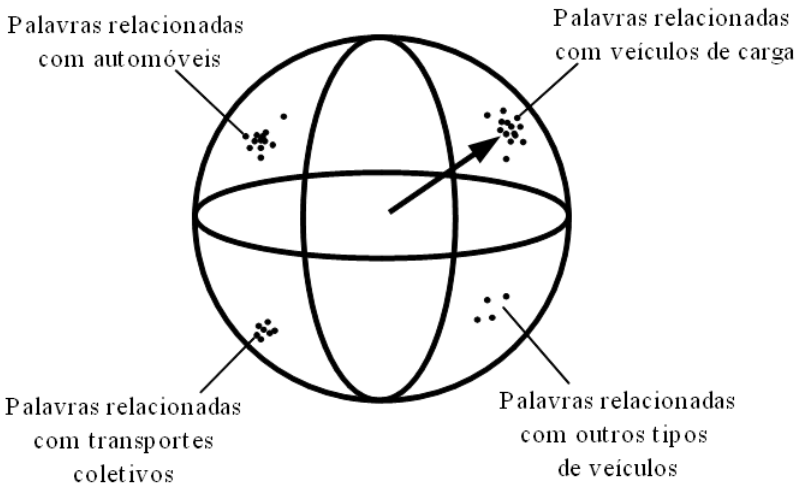

Figura 2.4: Vetores de termos

Adaptado de [Caid and Carleto, 2003] p. 6

A classificação dos documentos é feita por comparação entre os vetores que representam os documentos, agrupando os vetores que estão em uma mesma vizinhança no espaço n-dimensional. A Figura 2.4 ilustra uma hipotética distribuição

de vetores por assuntos em um hiperesfera, nela os vetores sobre um mesmo assunto apontam para uma mesma região. Como a proximidade entre dois vetores está sendo media pelo cosseno do ângulo entre eles, para vetores paralelos ou coincidentes o cosseno será máximo e para vetores ortogonais, o cosseno será zero, nesse caso os vetores são independentes entre si.

Os vetores são representados somente nos semi-eixos positivos do hiperespaço, isso evita que os vetores se tornem ortogonais apenas por possuírem um atributo antagônico. Por exemplo, “carro veloz” e “carro lento”, no lado (a) da figura 2.5, possuem similaridade no termo “carro”, mas são antagônicos nos termos “veloz” e “lento” e o cosseno do ângulo entre eles é igual a zero, ou seja não há similaridade entre eles. Por outro lado, se a representação for limitada aos semi-eixos positivos, conforme consta no lado (b) da figura 2.5, “carro veloz” e “carro lento” serão tomados com algum grau de similaridade, pois o cosseno do ângulo entre os vetores não será mais igual a zero, mesmo com a presença de termos que são opostos entre si [Polyvyanyy and Kuropka, 2007].

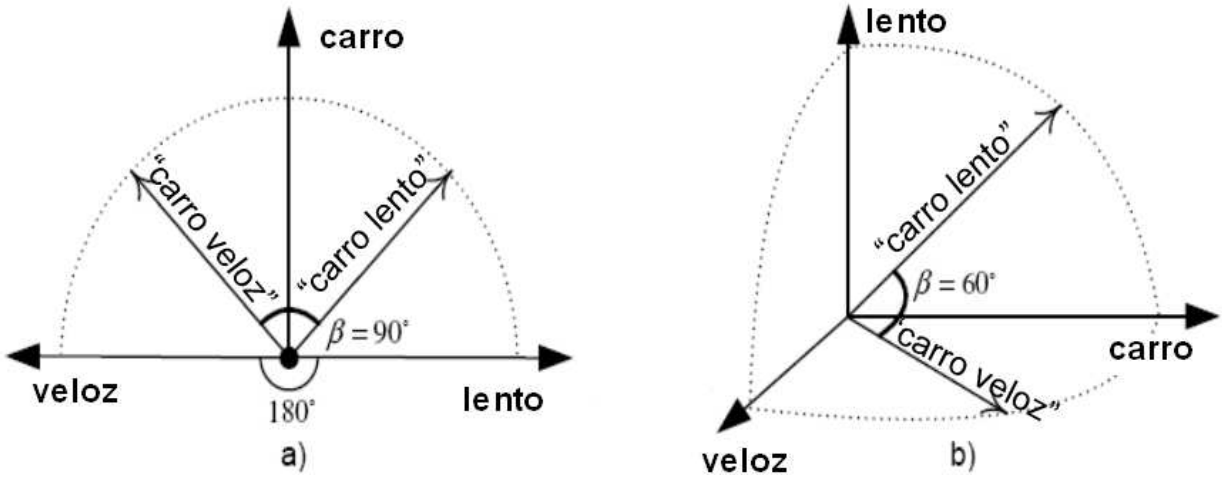

Figura 2.5: Vetores nos semi-eixos positivo

Fonte: [Polyvyanyy and Kuropka, 2007] p. 10

Admitindo que

$$\vec{d_1} = (p_{1,1}, p_{1,2}, \ldots, p_{1,n}) \text{ e}$$
$$\vec{d_2} = (p_{2,1}, p_{2,2}, \ldots, p_{2,n})$$

sejam representações vetoriais de dois documentos d_1 e d_2, onde $p_{i,j}$ é o peso do j-ésimo termo da base do espaço vetorial de termos na representação do i-ésimo documento.

Se $p_{i,j} = 0$, o termo em questão não consta no documento ou é irrelevante para sua representação. O critério de escolha dos termos a serem considerados na construção do espaço vetorial e a forma de calcular $p_{i,j}$ varia de acordo com a técnica empregada e são fatores que influenciam fortemente as medidas de desempenho do modelo.

A fórmula 2.5 expressa o cálculo da similaridade (Sim) entre dois documentos. Nela, $\vec{d_i}$ é uma representação vetorial do documento d_i.

$$Sim(d_1, d_2) = \cos\theta = \frac{p_{1,1}p_{2,1} + p_{1,2}p_{2,2} + \ldots + p_{1,n}p_{2,n}}{\sqrt{p_{1,1}^2 + p_{1,2}^2 + \ldots + p_{1,n}^2}\sqrt{p_{2,1}^2 + p_{2,1}^2 + \ldots + p_{2,n}^2}} \tag{2.5}$$

O uso de VSM permite a construção de classes para agrupamento automático dos documentos, todavia essas classes não são estanques, pois os vetores possuem uma distribuição no espaço n-dimensional e a partir de um documento é possível determinar os outros documentos que estão em sua vizinhança. Assim, se dois documentos são sabidamente pertencentes ás classes A e B, os documentos entre ambos possuem um grau de pertinência que muda gradativamente de uma classe para outra na medida em que os documentos são percorridos. Entretanto, a definição das fronteiras das classes torna-se algo vago e difícil de determinar.

Observa-se que os documentos tendem a concentrar-se em certas regiões do espaço n-dimensional. Esses agrupamentos podem ser utilizados para definir o assunto comum aos documentos da classe. A acurácia desse agrupamento depende

fortemente do poder que os termos que compõe a base do espaço vetorial têm de representar os documentos do *corpus*.

A operação de busca de documentos representados no modelo de espaço vetorial pode ser dividida em três etapa: (1) extração dos termos relevantes para representar o documento a ser procurado no espaço vetorial; (2) são atribuição pesos aos termos, usando a mesma métrica adotada para representação dos documentos no espaço vetorial e (3) cálculo da similaridade entre os vetores que representam o argumento de busca e os documentos armazenados, recuperando aqueles com maiores índices de similaridade.

2.2.1 WVSM - *Word Vector Space Model*

O WVSM utiliza diretamente as palavras do texto na construção dos vetores, desconsiderando somente as palavras que não carregam significado e que, por isso, não contribuem para caracterização do documento. Essas palavras descartadas são conhecidas como *stop words*. Essa abordagem produz vetores de grandes dimensões, impactando no desempenho das consultas, pois o crescimento da dimensão do espaço vetorial aumenta o consumo de espaço requerido para o armazenamento dos vetores que representam os documentos e incrementam o tempo necessário para realizar os cálculos envolvidos [Ikehara et al., 2001]. A falta de tratamento semântico produz distorções nos resultados, isto é documentos similares não são reconhecidos como tais (falso negativo) ou documentos não similares são recuperados como se fossem similares (falso positivo).

2.2.2 SVSM - *Semantic Vector Space Model*

O tratamento semântico dos termos utilizados para representação dos documentos leva a uma redução da dimensão do espaço vetorial e melhora a capacidade do modelo para recuperar documentos, mas aumenta o tempo necessário para construção

dos espaço vetorial em si e para análise das consultas no processo de conversão do argumento de busca em um vetor compatível com o espaço vetorial.

A escolha das palavras que representam os documentos e a determinação dos pesos dessas palavras na composição dos vetores do espaço vetorial podem ser feitas por meio de qualquer método estatístico para esse fim. Pode ser usado peso uniforme, por exemplo, um para as palavras representativas do documento e zero para as palavras que não ocorrem no documento ou são irrelevantes em sua representação vetorial. O peso mais usado é o TF-IDF (*term frequency, inverse document frequenct*) (fórmula 2.6).

Sejam t_i os termos de indexação, com $1 \leq i \leq m$, m dimensão do espaço vetorial, N o número de documentos presentes no *corpus* e n_i é o número de documentos nos quais o termo t_i ocorre.

$$idf_i = log(\frac{N}{n_i}) \tag{2.6}$$

Os pesos calculados pela fórmula 2.6, estabelecem uma relação inversa da capacidade de um termo em discriminar um documento dentro do *corpus*. Se um termo aparecer em todos os documentos ele não serve como discriminador e seu peso será zero [Oliveira et al., 2007], pois

$$log(\frac{N}{n_i}) = log(\frac{N}{N}) = log(1) = 0.$$

Oliveira e colaboradores (2007) expõem um experimento de classificação automática de um conjunto com 15 documentos do sítio extraídos do UOL e compara com uma classificação realizada por um classificador humano. Os documentos foram agrupados em 3 classes (economia, esporte e cinema), usando uma função de corte para restringir índices de semelhança a partir de um determinado patamar. Nesse experimento, Oliveira e seus colaboradores concluíram que não houve diferença entre a classificação manual e a automática utilizando o índice estatístico

TF-IDF.

Salton, Wong e Yang [Polyvyanyy and Kuropka, 2007] [Salton et al., 1975] propuseram uma combinação de fatores globais e locais no cálculo do índice TF-IDF (equação 2.7), onde w_{d,t_i} é o peso do termo t_i no documento d, α_{d,t_1} é o número de ocorrências do termo t_i no documento d, N é o número de documentos no *corpus* e n_{t_i} é o número de documentos no *corpus* onde o termo t_i aparece.

$$w_{d,t_i} = \frac{\alpha_{d,t_1}}{max_{t \in D} \alpha_{d,t}} \log(\frac{N}{n_{t_i}}) \tag{2.7}$$

Se um termo está presente em todos os documentos, ele não serve como discriminador e seu peso é zero, caso contrário seu peso é uma composição de seu peso relativo no documento com seu peso em todo o conjunto. Embora essa abordagem melhore o poder do discriminador, algumas limitações devem ser consideradas:

- documentos grandes são mal representados pois produzem vetores longos mas com produtos escalares pequenos;
- as palavras de um argumento de busca podem resultar em "falsos positivos", devido, por exemplo, a diferenças de inflexão dos termos;
- erros de digitação produzem resultados ruins na busca; e
- dificuldades, ou ausência de tratamento semântico reduzem os resultados recuperáveis de uma consulta, gerando "falsos negativos".

Entretanto, o método em si é simples, possui uma interpretação gráfica intuitiva, permite seu uso para classificação dos documentos e consultas *ad-hoc*. Variações desse método são largamente usadas, pois em sua forma original o custo para comparação dos vetores torna-se elevado devido ao crescimento da dimensão do espaço vetorial. O crescimento da base de representação implica em vetores esparsos, isto é com um elevado número de zeros nas coordenadas. Quanto mais diversificados forem os assuntos dos documentos que compõe o *corpus*, ou maiores

os documentos, mais esparsos serão os vetores. A inclusão e exclusão de documentos no *corpus* também têm custos elevados, pois alterações no conjunto de termos da base pode requerer ajustes nos vetores que representam cada documento do *corpus*. No caso de uso de pesos uniformes para os termos, esse impacto é mínimo, ao passo que no caso de se usar TF-IDF, ou outra forma de ponderação, a adição ou remoção de termos relevantes na base do espaço implicará em alteração nos pesos de todos os vetores do espaço vetorial.

Para contornar o crescimento da dimensão da base do espaço vetorial, Borko e Bernic [Borko and Bernick, 1963], em 1963, desenvolveram uma técnica denominada *KL method*. Por esse método a dimensão do espaço é reduzida pela escolha de novas bases considerando-se a semelhança semântica entre os termos da base. Outra abordagem para redução da dimensão do espaço de termos, é o uso da semântica latente LSI (*Latent Semantic Indexing*) que tenta encontrar novos significados por detrás do plural das palavras utilizadas na base do espaço vetorial, essa técnica permite reduzir a dimensão do espaço sem reduzir a qualidade da recuperação sobre os vetores [Ikehara et al., 2001].

O tratamento semântico dos termos que são usados para representar os documentos produz um espaço vetorial de dimensões menores, com isso termos com mesmo significado são agrupados em um único termo que atua como representante de uma classe de termos de significados semelhantes, isso pode ser realizado pela utilização de um tesauro [Ikehara et al., 2001].

Para construção da representação semântica é necessário levar em conta uma grande diversidade de fatores semânticos relacionados com a língua na qual os documentos estão expressos. Os seguintes fenômenos linguísticos podem ser destacados [Polyvyanyy and Kuropka, 2007]:

- **sinônimos** - palavras com mesmo significado ou com significados semelhantes que podem ser substituídas umas pelas outras;

- **inflexão** - são alterações nas palavras que refletem informações de tempo, gênero, quantidade e sujeito da ação;
- **composição** - quando duas ou mais palavras são justapostas formando uma nova palavra com um significado diferente, com em “guarda-chuva” ou “porta-bandeira”;
- **derivação** - processo de criação de outras palavras pela adição de um afixo à raiz da palavra, alterando também seu significado ou mudando sua categoria sintática, com por exemplo os prefixos “a” e “anti” (cromática $\rightarrow$ acromática, térmico $\rightarrow$ antitérmico), os quais funcionam como negação, ou o sufixo “mente”, que transforma verbos em adjetivos (rápido $\rightarrow$ rapidamente);
- **hiponímia** - palavras que representam instâncias de palavras que representam conceitos mais gerais, como por exemplo “vermelho”, “amarelo” e “azul” são hiponímias para “cor”,
- **meronímia** - palavras que representam a relação “parte de” ou “membro de”, por exemplo “motor”, é um meronímio para “carro” e “senador” é um meronímio de “político”;
- **homografia** - palavras com mesma grafia, mas significados diferentes, reconhecíveis pelo contexto em que ocorrem, por exemplo “manga” (de camisa ou fruta);
- **metonímia** - quando uma palavra é usada para representar outra associada a ela, como por exemplo “Ler Machado de Assis”, nesse caso “Machado de Assis” está no lugar de sua obra literária;
- **grupo de palavras** - são palavras que possuem significado próprio isoladamente, mas se juntam para formar uma expressão ou nome com significado diverso, por exemplo “São Paulo” possui significado diversos em cada ocorrência na frase “o apóstolo São Paulo nunca visitou a cidade de São Paulo nem

mesmo esteve no estado de São Paulo". Observe que a expressão "São Paulo", por si só forma um grupo de palavras mas seus significados na frase estão determinados por uma terceira palavra ("apóstolo", "cidade" ou "estado").

Além destes, há outros fenômenos que devem ser considerados, tais como: estrangeirismos, gírias, regionalismos e linguagem figurada, entre outros. Todos esses fenômenos linguísticos afetam grandemente os sistemas que pretendem utilizar informação semântica de forma implícita ou explícita.

Polyvyanny e Kuropka (2007) apresentam duas variações do modelo VSM: TVSM (*Topic Vector Space Model*)e eTVSM (*Enhanced Topic Vector Space Model*) os quais apresentamos a seguir.

2.2.3 TVSM - *Topic Vector Space Model*

Este método não pressupõe independência dos termos usados para indexar os documentos, por isso é mais flexível na determinação das similaridades. A base do espaço de representação dos documentos é composta por vetores de tópicos fundamentais como na figura 2.6. Os vetores da base são ortogonais e independentes entre si, restritos aos semi-eixos positivos de forma que, dados dois vetores quaisquer nesse espaço de representação, o ângulo entre eles estará entre 0^o e 90^o. Recio-Garcia e seus colaboradores [Recio-García et al., 2008] sugerem que um agrupamento (*clustering*) hierárquico ajudaria sistemas baseados em vetores a fornecer uma justificativa, mesmo que não muito clara, do motivo pelo qual um determinado registro é selecionado pelo mecanismo de recuperação de informação. Esse conceito é muito semelhante ao conceito de tópicos apresentados por Polyvyanny e Kuropka (2007).

Os tópicos são extraídos dos documentos através de heurísticas, isto é, não há um procedimento formal geral definido para tanto. Em todo caso, esse procedimento deve levar em conta os fenômenos linguísticos já mencionados: sinonímia,

inflexão, composição, derivação, hiponímia, meronímia, homografia, metonímia e grupos de palavras. Os tópicos serão representados por vetores ortogonais normalizados, isto é de comprimento unitário [Polyvyanyy and Kuropka, 2007].

Assim, todo termo $T_i \in T$, onde T é o conjunto de todos os termos, é expresso 3 dos vetores da base de tópicos, $\vec{t_i} = (t_{i,1}, t_{i,2}, \ldots, t_{i,k})$ de tal forma que sua direção indica a relevância do termo em relação aos tópicos que compõe a base e seu comprimento indica seu peso (importância), o qual deve estar no intervalo $[0, 1]$.

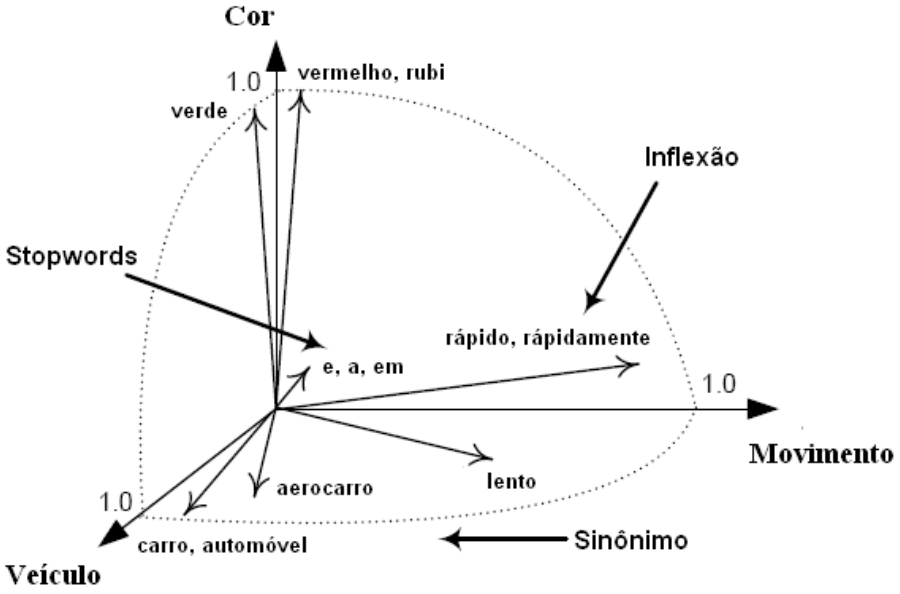

Figura 2.6: Vetores de tópicos no TVSM

Fonte: [Polyvyanyy and Kuropka, 2007] p. 11

Todo documento d_j do *corpus* é representado por um vetor normalizado $\vec{\delta_j}$, conforme a fórmula 2.8, na qual w_{d_j,t_i} representa o peso do tópico t_i no documento d_j. A determinação dos pesos dos tópicos pode ser feita da mesma forma que no VSM.

$$\vec{\delta_j} = \frac{\vec{d_j}}{|\vec{d_j}|} = \sum_{t_i \in T} w_{d_j,t_i} \vec{t_i} \tag{2.8}$$

Nesse modelo, sinônimos e inflexão de termos são representados por vetores paralelos. Composição com baixa interdependência dos termos são representados

através vetores cujos ângulos entre si tendem a 90^o, os ângulos entre os termos derivados com alta dependência entre si são representados por vetores cujos ângulos que tendem a 0^o. Hiponímias são representadas com ângulos pequenos e meronímias com ângulos que dependem dos níveis de agrupamento entre os objetos envolvidos.

Não existe ainda um mecanismo formal para determinar o comprimento dos vetores que representam os termos, nem o ângulo entre vetores inter-relacionados. Em geral recomenda-se usar comprimento um para termos relevantes (*content bearing words*) e zero para os demais termos, ou que se use um comprimento inversamente proporcional à frequência dos termos no documento, como acontece no TF-IDF.

2.2.4 eTVSM - *Enhanced Topic Vector Space Model*

Este método foi proposto originalmente por Kuropka (2003) com o objetivo é suprir as principais deficiências do TVSM: ausência de uma abordagem formal para determinar os comprimentos dos vetores, os ângulos entre vetores de termos inter-relacionados e o tratamento dos fenômenos linguísticos homografia, metonímia e grupo de palavras. Para isso, a similaridade dos documentos é calculada em função da similaridade do significado dos termos e não com base na similaridade dos termos em si através de interpretações [Polyvyanyy and Kuropka, 2007].

O modelo operacional do eTVSM utiliza-se dos conceitos: palavra, lema, termo, interpretação, e tópicos. As relações entre os termos são organizados em uma ontologia responsável por capturar informações acerca das relações entre os diversos conceitos presentes nos domínio dos conteúdos dos documentos do *corpus* [Kuropka, 2003].

Ontologia na ciência da computação pode ser definida como "*um modelo de dados (estrutura de dados) que representa um domínio e é usado para raciocinar acerca dos objetos daquele domínio e das relações entre eles*" [Florid, 2003].

Esse modelo é usado para raciocinar sobre os objetos do domínio e as relações entre eles. Para construção da ontologia representando as relações entre os termos, eTVSM utiliza os conceitos: termos, interpretações e tópicos. Esses conceitos são organizados hierarquicamente em um grafo orientado, não cíclico, no qual as arestas representam conceitos da mesma classe ou de classes inter-relacionadas. Essa hierarquia define associações entre os tópicos, que atuam como sub-tópicos e super-tópicos. As relações entre os tópicos são livres e podem ser de qualquer tipo, por exemplo: "parte de", "compõe" ou "é um". Um super-tópico pode possuir um número arbitrário de sub-tópicos. Um sub-tópico pode possuir um número arbitrário de super-tópicos. A única restrição para a estrutura é que ela deve estar livre de ciclos. O ângulo entre os vetores que representam os tópicos é determinado em função do grau de similaridade entre eles e não precisam ser ortogonais entre si [Polyvyanyy and Kuropka, 2007].

A estrutura do grafo não é necessariamente conexa, podendo possuir sub-grafos próprios ou tópicos isolados uns dos outros (desconexos). Se dois tópicos estão em sub-grafos desconexos distintos, eles são independentes entre si, logo os vetores que os representam são ortogonais.

Nessa abordagem, os vetores da base são tópicos, mas não necessitam ser ortogonais como no TVSM. O ângulo entre os vetores representa o nível de relação entre os tópicos na ontologia.

A determinação da similaridade entre os tópicos é construída em duas etapas: na primeira, utiliza-se um formalismo proposto por Kuropka (2003) para obter o mapa de tópicos. Na segunda etapa, calcula-se o produto escalar dos vetores que representam os tópicos. A Figura 2.7 representa uma estrutura de tópicos hipotética, na qual as setas indicam os sentidos dos relacionamentos entre os tópicos envolvidos, de forma que os tópicos mais gerais localizam-se na parte superior [Polyvyanyy and Kuropka, 2007].

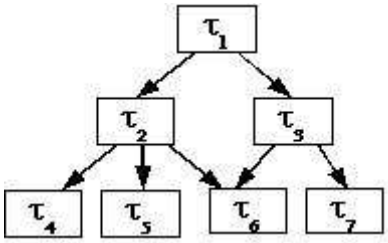

Figura 2.7: Hierarquia de tópicos

Fonte: [Polyvyanyy and Kuropka, 2007] p. 14

Seja $S(t)$ o conjunto dos super-tópicos do tópico t, então

$$S(\tau_1) = \{\}$$
$$S(\tau_2) = \{\tau_1\}$$
$$S(\tau_3) = \{\tau_1\}$$
$$S(\tau_4) = \{\tau_2\}$$
$$S(\tau_5) = \{\tau_2\}$$
$$S(\tau_6) = \{\tau_2, \tau_3\}$$
$$S(\tau_7) = \{\tau_3\}$$

Seja a relação de super-tópicos definida por

$$S^1(\tau_i) = S(\tau_i)$$
$$S^p(\tau_i) = \cup_{\tau_k \in S^{p-1}(\tau_i)} S(\tau_k), p > 1$$

S^* é denominado fecho transitivo de S e é dado por

$$S^*(\tau_1) = S^1(\tau_1) \cup S^2(\tau_1) \cup S^3(\tau_1) \cup \ldots$$

O conjunto θ de tópicos é divido em dois conjuntos disjuntos θ_N, o conjunto dos supertópicos e θ_L, o conjunto dos tópicos que não possuem subt-ópicos, isto é são folhas.

Um tópico T_i é representado por um vetor $\vec{\tau_i} = (\tau^*_{i,1}, \tau^*_{i,2}, \ldots, \tau^*_{i,t}) \in \Re^t$. A forma de cálculo dos vetores depende de eles serem folhas ou super-tópicos:

$$\forall \tau_i \in \theta_L : \vec{\tau_i} = (\tau^*_{i,1}, \tau^*_{i,2}, \ldots, \tau^*_{i,t})$$

onde

$$\tau^*_{i,k} = \begin{cases} 1, \text{ se } \tau^k \in S^*(\tau_i) \vee i = k \\ 0, \text{ caso contrário} \end{cases}$$

e

$$\forall \tau_i \in \theta_N : \vec{\tau_i} = \sum_{\tau_S \in \theta : \tau_i \in S(\tau_i)} (\tau_s)$$

Após a construção dos vetores eles são convertidos para a norma unitária, pois o que importa nesse modelo é a direção de cada vetor e não seu comprimento.

A ideia por trás dessa forma de cálculo é que, os tópicos que não possuem subtópicos funcionam com blocos básicos para construir os super-tópicos.

As interpretações são usadas para associar os termos aos tópicos e não podem estar associadas entre si. A cada interpretação ϕ pertencente ao conjunto de todas as interpretações Φ associa-se um peso $g(\Phi)$.

O vetor interpretação $\vec{\phi_i} = (\phi_{i,1}, \phi_{i,2}, \ldots, \phi_{i,t})$ pode ser definido como o vetor normalizado.

$$\vec{\phi}_{i,1} = \frac{g(\phi_i)}{|\sum_{\tau_k \in T(\phi_i)} \vec{\tau_k}|} \sum_{\tau_k \in T(\phi_i)} \vec{\tau_k}, \text{ onde } T(\phi_i) \in 2^\theta$$

Os termos são a menor unidade de informação à qual pode-se atribuir uma interpretação. Um termo pode ter múltiplas interpretações associadas. Um subconjunto especial dos termos é usado para resolver problemas de ambiguidade. Esse subconjunto é denominado termos de suporte e basicamente são termos que co-ocorrem no documento. Assim se um termo está associado a mais de uma interpretação, as co-ocorrências são usadas para identificar a interpretação mais adequada.

Um ontologia eTVSM é construída usando termos, tópicos (τ) e interpretações (ϕ). A Figura 2.8 ilustra uma ontologia.

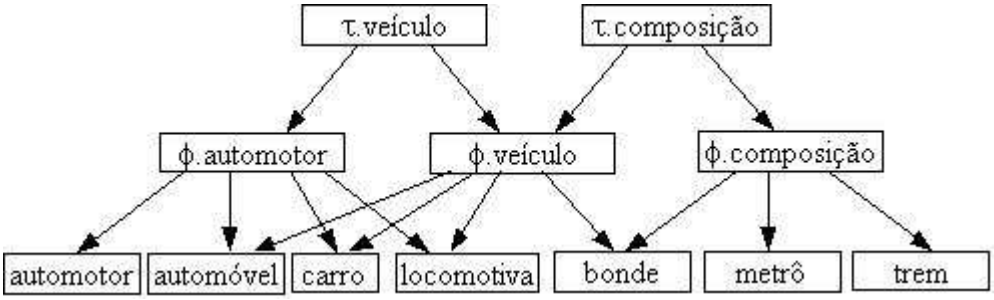

Figura 2.8: Exemplo de ontologia eTVSM

Fonte: Adaptado de [Polyvyanyy and Kuropka, 2007] p. 20

A tabela 2.2 apresenta um quadro comparativo das principais características das abordagens baseadas em espaços vetoriais.

Característica	**W-VSM**	**S-VSM**	**TVSM**	**eTVSM**
Complexidade	Baixo	Baixo-Médio	Médio-Superior	Superior
Consumo de memória	Elevado	Moderado	Moderado	Moderado
Dimensão do espaço vetorial	Elevada	Moderada	Moderada	Moderado
Poder de expressão semântica	Ausente	Moderado	Moderado-Superior	Superior
Capacidade para resolver ambiguidade nas consultas realizadas	Não	Não	Não	Sim
Uaa *stop list*	Sim	Sim	Sim	Sim
Necessita de um dicionário ou um tesauro, ou equivalente	Não	Sim	Sim	Sim
Construção automática	Sim	Sim	Não	Parcial

Tabela 2.2: Comparação das abordagens de RI baseadas em espaço vetorial

2.3 Avaliação dos Sistemas de Recomendação

A qualidade de um sistema de recomendação depende de diversos fatores como tempo de resposta, quantidade espaço utilizado para armazenamento dos dados,

número de resultados apresentados e número de sugestões relevantes para o usuário do sistema. Uma recomendação somente é útil se for considerada relevante por quem a recebe, ou seja em função de sua utilidade. Infelizmente isso não é algo simples de ser avaliado, pois, como já foi mencionado anteriormente, a relevância é um critério pessoal, não podendo ser medido diretamente, a não ser que seja explicitado pelo próprio usuário. As duas medidas principais da qualidade de um sistema de recomendação são *precision* (fórmula 2.9) e *recall* (fórmula 2.10), além destas, existem outros índices como por exemplo *fallout* (2.11) e *F-measure* (fórmula 2.12), também conhecido como F_1 [van Rijsbergen B, 1979], cujas fórmulas são apresentadas a seguir:

	RELEVANTE	NÃO RELEVANTE	
RECUPERADO	$A \cap B$	$\bar{A} \cap B$	B
NÃO RECUPERADO	$A \cap \bar{B}$	$\bar{A} \cap \bar{B}$	$\bar{B}$
	A	$\bar{A}$	N

Tabela 2.3: Tabela de contingência

Fonte: [van Rijsbergen B, 1979] p 114

Onde N é o número de documentos no *corpus*.

Precision fornece uma estimativa da probabilidade condicional de um item ser recuperado dado que ele é relevante.

$$\textit{Precision} = \frac{|A \cap B|}{|B|} \tag{2.9}$$

Recall fornece uma estimativa da probabilidade condicional de um item ser relevante dado que ele foi recuperado.

$$\textit{Recall} = \frac{|A \cap B|}{|A|} \tag{2.10}$$

Fallout fornece uma estimativa da probabilidade condicional de um item ser recuperado dado que ele é não relevante.

$$\textit{Fallout} = \frac{|\bar{A} \cap B|}{|\bar{A}|} \tag{2.11}$$

F-measure é a média harmônica entre *recall* e *precision*.

$$F_1 = \frac{2}{\frac{1}{\textit{recall}} + \frac{1}{\textit{precision}}} = \frac{2.\textit{precision}.\textit{recall}}{\textit{precision} + \text{recall}} \tag{2.12}$$

No contexto de recuperação de informação, os documentos relevantes são aqueles que satisfazem a intenção da busca. Isso pode ser substancialmente diferente do resultado da busca, pois não se trata de uma busca exata, cujos resultados podem conter falsos positivos ou omitir documentos que eram esperados no resultado da consulta, mas que não foram localizados pelo mecanismo de determinação de similaridade.

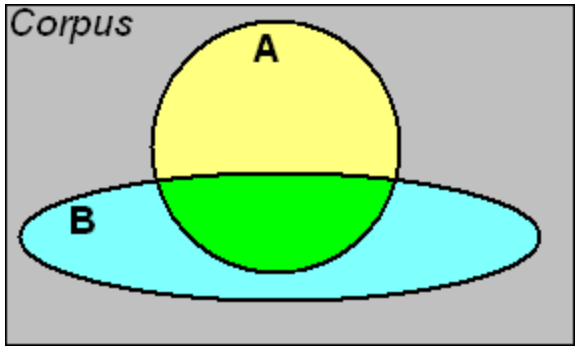

Figura 2.9: A=documentos relevantes e B=documentos recuperados

N=Número de documentos no *corpus*

O diagrama de Venn (figura 2.9) ilustra a situação. Se o conjunto de documentos recuperados B crescer até que todos os documentos do *corpus* sejam recuperados, o índice *precision* poderá crescer até que todos os documentos relevantes sejam recuperados, decaindo daí para frente até atingir seu valor mínimo $\frac{|A|}{N}$. Por outro lado, *recall*, *fallout* e crescerão até 1, que é o valor máximo para ambos. Para uma recuperação de 100% com 100% de precisão, teríamos $A = B$.

Capítulo 3

Exemplos de sistemas de recomendação

Um dos mais famosos sistema de recomendações de comércio eletrônico é o da Amazon.com^{TM}, utilizando uma estratégia que recomenda para um determinado consumidor produtos comprados por outros consumidores com perfis semelhantes [Shahabi and Chen, 2003]. Os sites de comércio eletrônico também implementam recomendações baseadas na navegação e nas compras dos clientes, trazendo resultados do tipo: "quem consultou X, também consultou Y", "quem comprou X, também comprou Y", além de oferecer produtos baseados nas similaridades entre eles ou em associações que consideram que a utilidade do produto X para quem adquiriu o produto Y, por exemplo, ao comprar um notebook, provavelmente vou precisar de um roteador sem fio, ou de uma mochila para transportá-lo.

Nas seções seguintes, serão apresentados sistemas de recomendação propostos como resultados de projetos de pesquisa ao invés de sistemas comerciais. Dois dos sistemas escolhidos possuem relação direta com o Currículo Lattes: o Sistema de Recomendação de Bibliotecas Digitais e uma proposta de uso de recomendação para recuperação de perfis de pesquisadores a partir do Currículo Lattes.

3.1 Sistema Yoda

Yoda é um sistema de comércio eletrônico proposto por Chen e Shahabi (2003), esse sistema utiliza uma abordagem híbrida de dois passos combinando filtragem colaborativa e filtragem baseada em conteúdo (figura 3.1): primeiro um processamento em lote, no qual são geradas listas de classes de recomendação baseada no comportamento do usuário enquanto navega na rede, combinado com técnicas de análise de conteúdo. Esse sistema mantém uma lista de recomendações elaboradas por especialistas humanos e classes representativas das notas atribuídas pelos usuários às recomendações recebidas [Shahabi and Chen, 2003].

No segundo passo, o sistema utiliza informações sobre a navegação do usuário na utilização do sistema. Com base nessa informação, Yoda estima o grau de confiança que os usuários exibem nas recomendações realizadas pelos especialistas humanos no sistema, produzindo recomendações cruzadas entre os perfis utilizando o grau de confiança estimado como pesos para as novas recomendações.

Para reduzir a complexidade computacional e o tempo de processamento, Yoda aplica uma otimização de agregação *fuzzy*. O sistema incorpora um módulo de aprendizagem baseado em algoritmos genéticos para ajustar a confiança do usuário na recomendação, analisando somente o histórico de navegação do usuário, sem necessidade de preenchimento de questionários, perfis ou recomendações por parte do usuário.

3.2 Sistema *Implicit*

Birukov e seus colaboradores (2005) propuseram esse sistema para uso em pequenas comunidades de usuários, seu objetivo é realizar recomendações baseado colaboração implícita a partir da análise dos resultados das buscas submetidas pelos usuários e na navegação pelas páginas resultantes (figura 3.2). Usa agentes inteligentes para coletar informações sobre as consultas submetidas pelos usuários

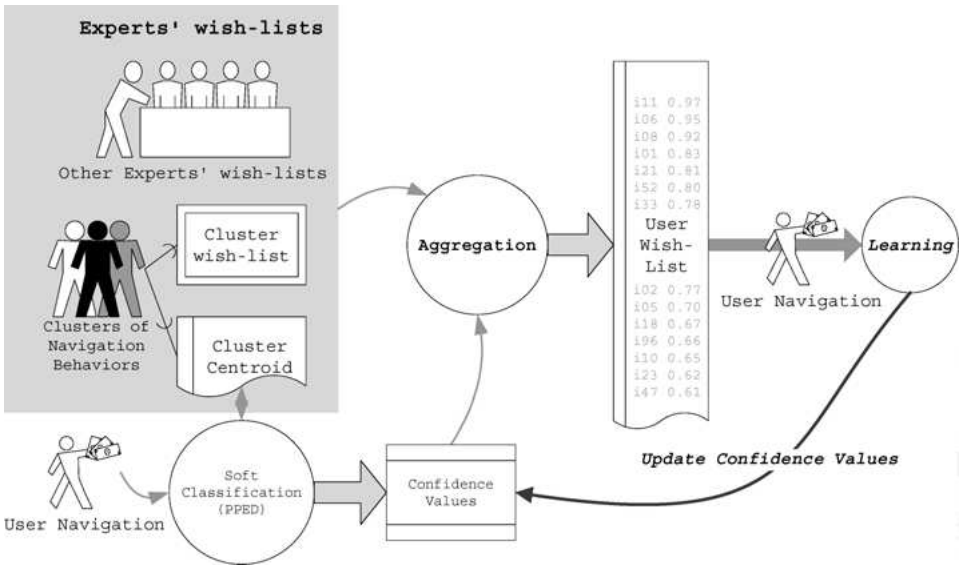

Figura 3.1: Fluxo de processo do sistema Yoda

Fonte: [Shahabi and Chen, 2003] p. 177

a um mecanismo de busca externo e compõe recomendações colaborativas implícitas. Cada agente é responsável por coletar informações sobre o comportamento de um determinado usuário enquanto este estiver conectado no sistema; usar o histórico de navegação do usuário para fornecer sua melhor recomendação para outros agentes, considerando a similaridade entre os argumentos de busca submetidos por cada usuário e realizar recomendações para o usuário que ele atende. Para realizar as recomendações ao seu usuário, cada agente combina os resultados das consultas submetidas ao buscador externo com as recomendações recebidas dos outros agentes [Birukov et al., 2005].

O sistema *Implicit* não requer nenhuma instalação do lado do usuário e utiliza do lado do servidor o padrão JADE (*Java Agent Development Framework*) para desenvolvimento de agentes.

Em testes controlados essa metodologia produziu um crescimento dos índices *precision* e *recall* com o crescimento do número de agentes utilizados.

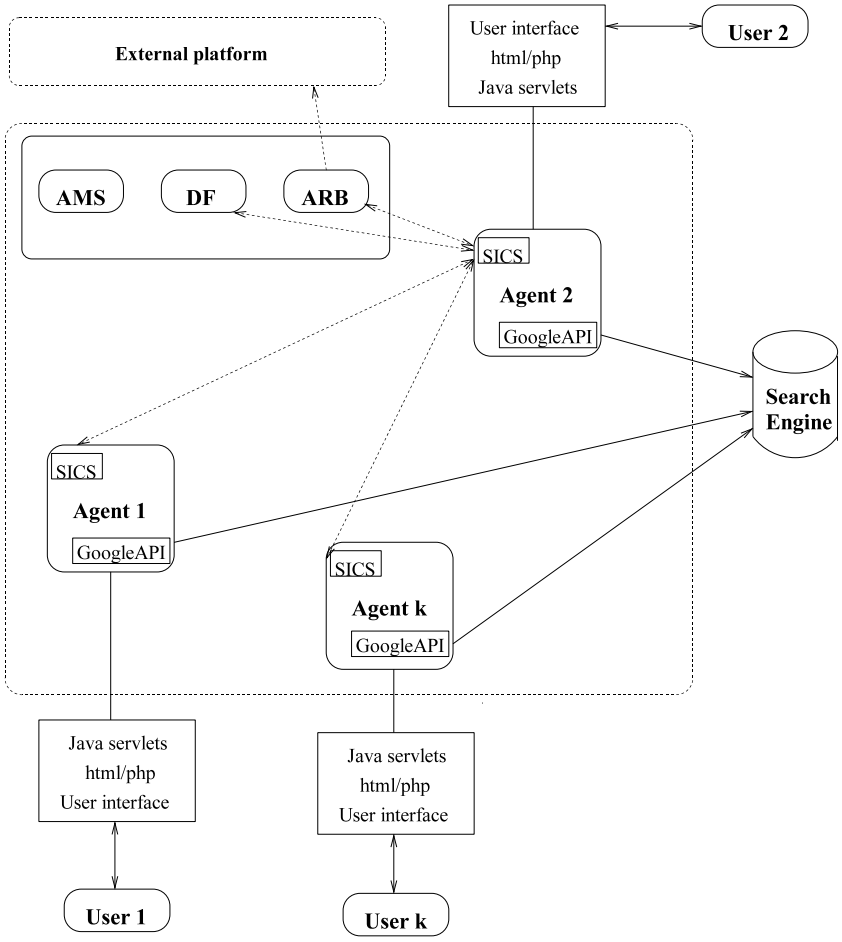

Figura 3.2: Arquitetura do sistema *Implicit*

Fonte: [Birukov et al., 2005]

3.3 Sistema W-RECMAS

O sistema W-RECMAS (*Recommender System to Web based on Multi-Agent System for academic paper recommendation*) é um sistema para recomendações de publicações acadêmicas e comunidades virtuais de aprendizagem cujo objetivo é auxiliar o os usuários a trocarem informações e compartilharem conhecimento entre os membros da comunidade [Cazella and Alvares, 2005]. W-RECMAS foi proposto por Cazella e Alvares (2005), sua ideia principal baseia-se no comportamento humano de procurar colegas que tenham mais conhecimento sobre determinado assunto para obter opiniões relevantes sobre algum assunto de interesse.

Trata-se de um sistema híbrido que combina avaliação das recomendações produzidas pelo sistema com a análise dos perfis do usuários e das comunidades que ele participa. A análise dos perfis usa técnicas de *data mining* por meio multiagentes. W-RECMAS utiliza informações da comunidade virtual do usuário, informações do perfil do usuário combinadas com informações extraídas do conjunto de perfis dos usuários pela aplicação de técnicas de regras de associação.

Esse sistema é responsável pela criação e recomendação de comunidades acadêmicas virtuais e seus agentes possuem habilidades e comportamentos diversos e cada um deles é responsável por uma tarefa específica.

A Figura 3.3 exibe a arquitetura do sistema W-RECMAS. Nela pode-se ver a diversidade de agentes presentes no modelo. Os agentes do tipo *crawler* localizam-se no servidor e são responsáveis por obter os Currículos Lattes dos usuários e mantê-los atualizados no sistema. Os agentes do tipo *personal* rodam nas máquinas dos usuários e são responsáveis por apresentar aos usuários as recomendações do sistema e observar o comportamento dos usuários, fornecendo retroalimentação para o sistema. Os agentes do tipo *recommender* enviam textos de recomendações com as devidas explicações para os agentes do tipo *personal*. Os agentes do tipo *community* estabelecem as comunidades e identificam usuários potenciais a serem recomendados para ingressar na comunidade. Os agentes do tipo *analyst* possuem um conjunto variado de responsabilidades: analisar os currículos, calcular os índices de recomendação, encontrar as maiores similaridades entre usuários e itens disponíveis e aplicar *data mining* para identificar novas áreas de interesse dos usuários.

3.4 Sistema de Recomendação para Bibliotecas Digitais

Lopes (2006) propôs uma metodologia para recomendação de publicações acadêmicas a partir de informações extraídas do Currículo Lattes de pesquisadores. Essa

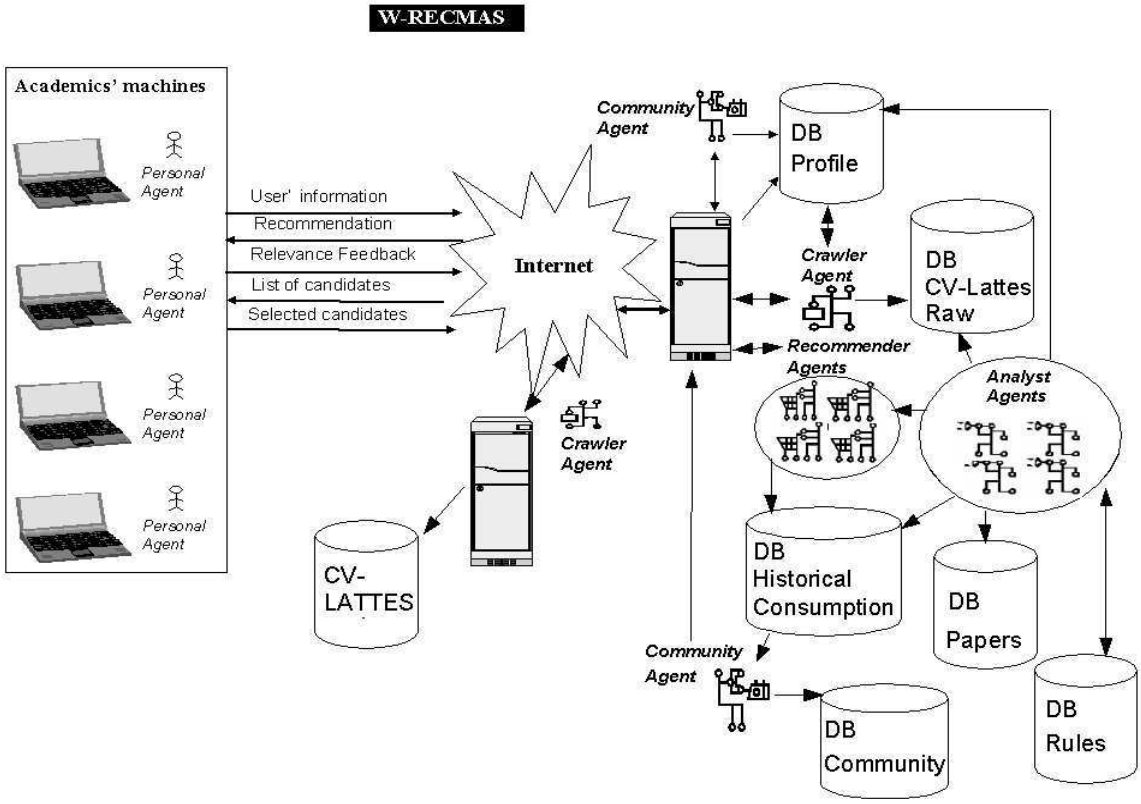

Figura 3.3: Arquitetura do sistema W-RECMAS

Fonte: [Cazella and Alvares, 2005]

metodologia foi implementada em um sistema piloto para recomendar publicações da área de Ciência da Computação.

A metodologia proposta pela autora utiliza como perfil do usuários informações extraídas do Currículo Lattes e informações descritivas das publicações através de metadados no formato Dublin Core (http://dublincore.org/). O sistema funciona como um provedor de serviços. A Figura 3.4 ilustra sua estrutura.

O sistema recolhe dados sobre as publicações, interpreta esses dados, extrai, cataloga e armazena as informações relevantes para representação das publicações bibliográficas. Os perfis dos usuários são extraídos dos respectivos Currículos Lattes submetidos ao sistema, no formato XML. O sistema representa os perfis dos usuários e as publicações acadêmicas através de um modelo de espaço vetorial.

São aplicadas técnicas próprias para redução da dimensão da base de vetores. As palavras-chave informadas pelos usuários são tomadas como descritores e inseridas integralmente nos vetores. O sistema utiliza um esquema de pesos para os termos, que depende: da localização do termo na produção bibliográfica; do idioma; da formação acadêmica e da produção bibliográfica mais recente. Esses pesos são combinados com a aplicação da técnica TF-IDF para determinar da importância relativa de um termo como descritor de um documento. Os resultados são apresentados aos usuários para que sejam avaliados. As avaliações são realizadas pela classificação dos resultados por meio de termos vagos: péssimo, ruim, médio, bom, ótimo e do próprio autor.

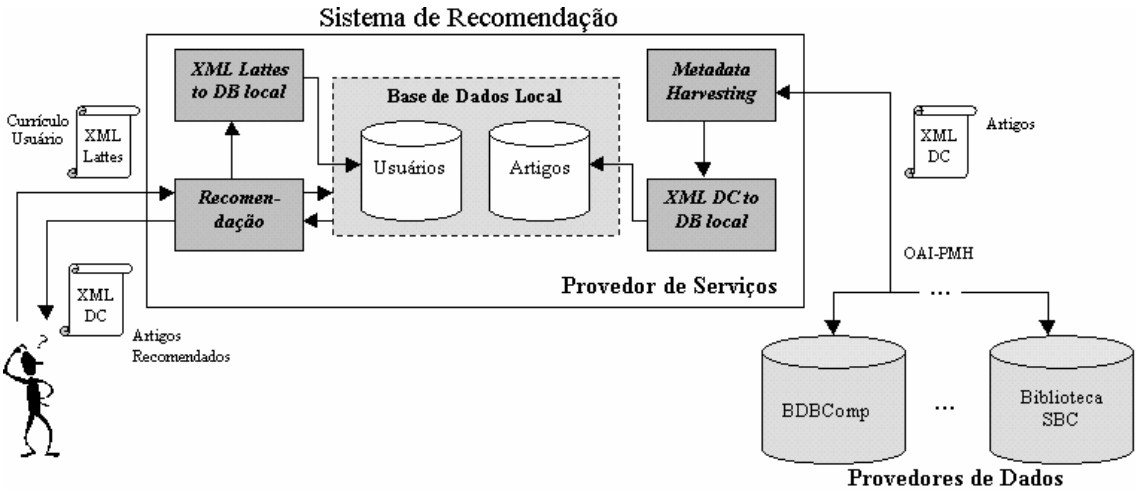

Figura 3.4: Modelo do Sistema de recomendação para Bibliotecas Digitais

Fonte: [Lopes et al., 2006] p. 37

3.5 Currículo Lattes – uso de recomendação para recuperação de perfis

Marques (2007), apresenta uma proposta de recomendação para recuperação de Currículo Lattes na qual sugere uma abordagem para recuperação de perfis de

usuários. A abordagem proposta foi testada em um site de relacionamento denominado Dois Corações com resultados considerados satisfatórios pelo autor.

A abordagem proposta por Marques é híbrida, utilizando dados extraídos do Currículo Lattes dos pesquisadores, dados coletados implícita e explicitamente das interações com os usuários, e regras controladas pelo administrador do sistema. Sugere a utilização de indicadores extraídos dos currículos, tais como área de atuação, tempo de experiência nas áreas de atuação e indicadores da produção científica dos últimos anos, como produção bibliográfica (livros, artigos, periódicos e capítulos de livros), orientações realizadas e em andamento e participação em eventos.

Entre as informações obtidas a partir da colaboração dos usuários constam:

- avaliações realizadas mutuamente pelos pesquisadores;
- informações prestadas sobre semelhança entre perfis de pesquisadores pelos próprios pesquisadores;
- dados navegacionais coletados durante a utilização do sistema após o usuário ter se autenticado para realizar buscas.

Marques sugere a possibilidade de uso de buscas armazenadas a serem executadas periodicamente por robôs. Os resultados dessas buscas podem ser encaminhados para o usuário que as submeteu sempre que houver alguma inclusão de novos currículos no conjunto resultante da busca. O autor destaca a importância da temporalidade das recomendações e sugere que as consultas considerem todo o histórico curricular dos usuários durante uma busca, ou sejam restringidas a um período específico. Isso é particularmente útil quando se está diante de uma quebra de paradigma tecnológico e há poucos especialistas com conhecimento sobre o assunto, nesse caso é interessante reduzir o escopo dos perfis a um período que englobe a mudança tecnológica. Além disso, pesquisadores que não publicam mais, mas que já publicaram muito no passado seriam excluídos paulatinamente dos re-

sultados das buscas na medida em que fossem ficando fora do intervalo de tempo considerado, privilegiando currículos mais atualizados.

Outro conceito incluído na proposta metodológica é a relevância, isto é, o grau de importância ou confiabilidade de uma recomendação realizada por um pesquisador. A relevância é resultado de uma composição de diversos fatores que incluem critérios como: área de atuação - opiniões emitidas fora das áreas em que o pesquisador atua são menos relevantes; avaliações recebidas pelos seus pares - pesquisadores com melhor avaliação e que, portanto, gozam de melhor reputação no meio acadêmico devem ter opiniões mais confiáveis; produção científica - pesquisadores com produção científica maior e mais recente têm opiniões mais relevante.

Citando Cazella e Álvares (2005), Marques (2007) propõe o uso de um *ranking* de recomendação calculado pela composição dos escores dos indicadores. A fórmula 3.1 é utilizada para calcular o valor normalizado de a: $\bar{a}$.

$$\bar{a} = MinMax(a) = \frac{a - a_{min}}{a_{max} - a_{min}}(a_{new\ max} - a_{new\ min}) + a_{new\ min} \tag{3.1}$$

Por construção, $\bar{a}$ é tal que, $a_{new\ min} \leq \bar{a} \leq a_{new\ max}$, é o valor normalizado resultante se aplicado ao valor original $a_{min} \leq a \leq a_{max}$. Dessa forma, variáveis com valores em diferentes intervalos podem comparadas após conversão para um intervalo único.

As variáveis normalizadas podem ser utilizadas para o cálculo do *ranking* de recomendação (fórmula 3.2):

$$RR = \frac{\sum_{i=1}^{n} \overline{a_i} p_i}{\sum_{i=1}^{n} p_i} \tag{3.2}$$

Além dos critérios quantitativos expressos matematicamente, o autor sugere a utilização de dados qualitativos obtidos pela avaliação direta ou implícita das recomendações realizadas e da avaliação dos currículos pelos usuários.

O modelo do sistema proposto por Marques inclui os módulos de cadastro, indexação, consulta, recuperação e recomendação:

- Módulo de cadastro: entrada de dados do sistema.
- Módulo de indexação: identifica e indexa as informações relativas aos perfis dos usuários.
- Módulo de consulta: interface entre o banco de dados e o ambiente de execução. É responsável por interagir com o usuário para obter dados que possam ser enviados ao sistema com o fim de recuperar informação.
- Módulo de recuperação: recebe as informações de módulo de consulta e filtra o que é relevante.
- Módulo de recomendação: manipula os dados dos perfis dos usuários e avaliações dos itens com o objetivo de criar regras e extrair informações para composição das listas de recomendação.

A proposta apresentada foi implementada em um site de relacionamentos (2 Corações) e foi transposto, do ponto de vista teórico, para recuperação de perfis de usuários do Currículo Lattes, essa transposição não chegou a ser implementada.

Capítulo 4

Problema abordado

Este capítulo expõe a sistemática de recomendação de consultores *ad-hoc* em uso no CNPq, detalhes de seu funcionamento dentro do contexto da avaliação das propostas submetidas àquele Conselho, concluindo com uma análise do desempenho geral do sistema. Para fins de análise, será considerado sucesso a efetiva emissão de parecer por consultor indicado dentre os consultores recomendados pelo sistema. Consultores que não tenham emitido o parecer, ou que tenham sido dispensados da emissão do parecer, serão considerados como insucesso, e, da mesma forma, consultores indicados sem que tenham sido recomendados pelo sistema.

O sistema de recomendação de consultores em uso no CNPq foi implantado em no segundo semestre de 2006, é uma ferramenta de apoio ao trabalho dos técnicos do CNPq na tarefa de selecionar consultores para avaliação *ad-hoc* de propostas submetidas ao CNPq.

O corpo de consultores é registrado em um banco de consultores, o qual é composto por bolsistas de produtividade em pesquisa do CNPq, os quais têm a obrigação contratual de prestar consultoria ao CNPq atuando como pareceristas, e por outros pesquisadores de renome convidados a emitir parecer, atuando voluntariamente na avaliação das propostas.

4.1 Indicação de consultores no âmbito do CNPq

O CNPq fomenta o desenvolvimento científico e tecnológico através de ferramentas de apoio à pesquisa e à formação de recursos humanos através de instrumentos como: auxílio para realização de eventos, auxílio para participação em eventos, auxílio para editoração, financiamento de projetos de pesquisa e desenvolvimento, concessão de bolsas estudo no país e no exterior, etc.

Para ter acesso a esses recursos, os candidatos submetem propostas ao CNPq, essas propostas são agrupadas em editais e chamadas, conforme sejam regidas por editais públicos, ou por normas do CNPq. Editais, ou chamadas, podem ser criados para implementar convênios entre CNPq e outros órgãos. Por suas características, os editais são concorrências públicas com fins determinados ou genéricos. Como exemplos de editais com fins determinados, pode-se citar o apoio ao desenvolvimento de tecnologias para um setor de aplicação específico, como no caso dos fundos setoriais (CT-Energia, CT-Amazônia, CT-Petro, etc.) e, como exemplo de editais genéricos, os editais universais. Os editais servem também para concessão de bolsas de formação e de estudo, no Brasil e no exterior, ou ainda bolsas de incentivo à pesquisa. Por outro, lado as chamadas implementam políticas permanentes de financiamento, como participação em eventos, realização de congressos e concessão de bolsas estudos. Em todos os casos, há uma divisão do edital, ou da chamada, em períodos de submissão, resultando em julgamento e contratação em lotes. Todas as propostas de um mesmo período são julgadas em conjunto e concorrem entre si pelos recursos disponíveis.

Abstraindo o conceito de proposta e concentrando a atenção nas características gerais, uma proposta possui basicamente um proponente; zero ou mais membros adicionais na equipe do projeto e uma documentação de detalhamento do objeto da proposta. Essa documentação normalmente é materializada em um documento de projeto, ou proposta de trabalho, contendo objetivos, prazos, recursos, contra-

partidas, metodologia a ser utilizada, resultados esperados e outras informações relevantes para avaliação da proposta. Parte desses dados são registrados em estruturas relacionais, parte na forma de documentos textuais e ainda em arquivos digitais anexos à proposta.

Os currículos dos proponentes são considerados parte integrante da proposta e são usados na avaliação destas. Os dados curriculares dos são armazenados em tabelas relacionais e possuem uma imagem em formato XML. Alguns desses dados são textuais como por exemplo títulos de produções científicas, palavras-chave para indexação e nomes de coautores em produções científicas e tecnológicas.

Para que possa ser contratada, as propostas passam por um processo de avaliação composto por várias etapas ou fases. São ao todo quatro fases, ou etapas, sendo que a primeira e segunda podem ocorrer em paralelo:

Pré-seleção: a rigor, não se trata de um julgamento, mas de uma verificação se o proponente e o objeto da proposta estão de acordo com os requisitos estabelecidos pelo edital ou chamada ao qual foi submetido.

Parecer de consultor *ad-hoc*: consiste na avaliação de mérito científico e tecnológico da proposta realizada por especialistas nos domínios do conhecimento relacionados à proposta. O resultado dessa análise é uma recomendação de aceitação, ou rejeição, da proposta em função de seu mérito tecnológico e científico, da viabilidade de sua execução e de outros aspectos como inovação, relevância e capacidade da equipe de projeto para realizá-lo. O parecer dos consultores é subsidiário, não terminativo.

Avaliação por Comitê de Assessoramento (CA): os comitês de assessoramento são órgãos colegiados cujos membros são nomeados dentre listas de especialistas escolhidos por votação pelos seus pares, possuem mandato fixo e delegação para realizar análise de mérito científico e tecnológico das propostas. Os pareceres dos consultores *ad-hoc* são utilizados como subsídios para a avaliação realizada pelos comitês de assessoramento. O resultado final da análise dos comitês são,

a grosso modo, duas listas: uma com as propostas sem mérito para aprovação e, outra, com as propostas com mérito para aprovação em ordem de prioridade de atendimento.

Deliberação final por Diretoria: ratificação dos pareceres desfavoráveis exarados pelos comitês e aprovação final das propostas aprovadas pelos comitês respeitando a disponibilidade de recursos. As propostas com parecer favorável dos comitês são classificadas em uma lista única, de acordo com as prioridades estabelecidas por cada comitê. As propostas que estiverem dentro da disponibilidade orçamentária recebem parecer final de aprovação e são encaminhadas para contratação, ao passo que, aquelas que não alcançarem prioridade suficiente para atendimento, recebem parecer desfavorável, mas não de mérito.

A indicação de consultor, para avaliação de proposta, é apoiada por um sistema de recomendação automática. Esse sistema de recomendação extrai uma aproximação dos perfis dos consultores, dos proponentes e das propostas para comparação entre eles a partir de informações textuais como palavras-chave e títulos das propostas, e da produção científica dos pesquisadores envolvidos. Outras informações correntes no sistema como áreas de atuação, comitê de assessoramento de vínculo, instituição de vínculo empregatício, instituição de execução da proposta e outros são considerados no processo de recomendação.

A Figura 4.1 apresenta um diagrama que exibe o contexto do sistema de recomendação de consultores em uso no CNPq. As sugestões são realizadas por meio de heurística cujos parâmetros são configuráveis por edital/chamada e sua execução se dá de duas formas: execução em lote logo após o final do período de submissão das propostas e sugestão sob demanda, realizada quando o técnico rejeita as sugestões anteriores e solicita novas sugestões.

Atualmente há duas formas para indicar um consultor para avaliar uma proposta: uma auxiliada por um processo de recomendação automática realizada pelo sistema e a outra sem ajuda do sistema. Em ambos os casos o técnico faz uso de um

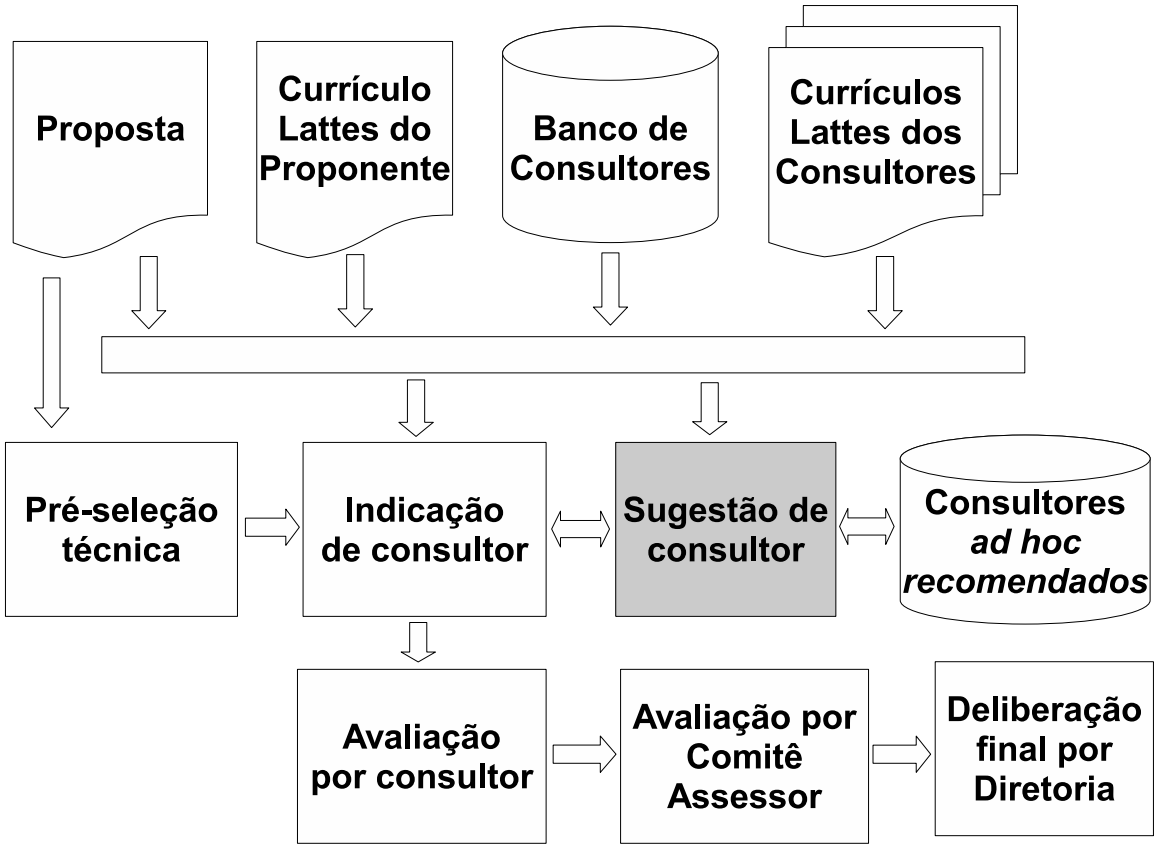

Figura 4.1: Diagrama de contexto da recomendação de consultor

banco de consultores o qual é composto por bolsistas de produtividade em pesquisa e por pesquisadores de destaque na comunidade científica. O técnico deve comparar os Currículos Lattes do proponente com os currículos dos consultores, além de considerar outras informações como:

- quem são os membros da equipe do projeto;
- a produção científica dos consultores e dos proponentes; e
- possíveis vínculos que possam produzir desvios que coloquem em dúvida a isenção dos consultores devido a conflito de interesses.

A indicação de consultores está sujeita a algumas limitações importantes como:

- pouco tempo para realizar a indicação de consultores para um grande número de propostas;

- dificuldade para os técnicos manterem atualizados seus conhecimentos sobre os perfis dos consultores;
- risco de acúmulo de indicações em alguns consultores mais conhecidos; e
- não utilização dos novos consultores habilitados.

Para ajudar no processo de indicação de consultores, o sistema de recomendação automática gera um conjunto de sugestões que são apresentadas aos técnicos no momento da indicação.

A recomendação automática pelo sistema usa abordagem baseada no conteúdo através de um índice de similaridade entre os perfis dos consultores habilitados e dos proponentes e entre consultores e propostas. Esse índice de similaridade é calculado pela contagem relativa das palavras extraídas das palavras-chave e títulos da produção científica presentes no Currículos Lattes dos proponentes comparadas com as mesmas informações dos currículos dos consultores. Da mesma forma, as palavras extraídas das palavras-chave e títulos das propostas são comparadas com as palavras-chave e títulos da produção científica presentes no Currículos Lattes dos consultores. Na construção dos perfis, são excluídas as palavras de pouco valor semântico (*stop words*) contidas em uma lista específica *stop list*. Convém ressaltar que essa contagem de palavras é realizada pelo *Oracle InterMedia Text* do SGDB Oracle^{TM} 10g. Essa ferramenta devolve índice relativo (*score*) de ocorrências dos termos pesquisados no conjunto de registros indexados. Esse índice determinado pelo próprio SGDBe varia de zero a cem, onde zero significando nenhuma ocorrência localizada até cem todas as ocorrências localizadas.

Cada consultor habilitado é cadastrado em um banco de consultores e possui um nível que corresponde a um nível de bolsa de produtividade em pesquisa do CNPq. Esse nível é comparado com o nível dos proponentes. Se o consultor não possuir bolsa de produtividade em pesquisa junto ao CNPq, um nível é atribuído a ele no ato de seu cadastro no banco de consultores. O proponente que não seja beneficiário

de bolsa de produtividade em pesquisa recebe o nível mais baixo de referência para efeitos da escolha do consultor.

O sistema considera ocorrências de coautoria em produção científica e tecnológica a partir das citações bibliográficas informadas pelos pesquisadores em seus currículos.

Os parâmetros usados no cálculo de similaridade para recomendação de consultores foram escolhidos empiricamente em um processo de prototipagem e testes, no qual foram geradas recomendações que foram avaliadas por gestores até serem consideradas satisfatórias. Os critérios são:

Critérios de similaridades positivos – aumentam a probabilidade de recomendação:

- especialidade da área do conhecimento: maior peso para maior aproximação entre as áreas de atuação do consultor e do coordenador do projeto;
- comitê de assessoramento: maior peso se o consultor e o coordenador do projeto forem ligados ao mesmo comitê de assessoramento que irá julgar o projeto na fase seguinte;
- similaridade dos perfis curriculares dos consultores e dos coordenadores de projeto: maior peso para maior similaridade;
- similaridade do perfil do consultor em relação ao projeto: maior peso para maior similaridade; e
- níveis do consultor e do coordenador do projeto: maior peso para consultores de nível mais alto.

Parâmetros de similaridades negativos – reduzem a probabilidade de recomendação:

- proximidade da instituição de vínculo: menor peso para maior aproximação entre consultor e coordenador do projeto e entre consultor e instituição de execução do projeto da proposta; e
- número de propostas para as quais o consultor já foi indicado: menor peso para consultores com mais propostas para avaliar dentro do edital/chamada.

Parâmetros de excludentes – impeditivos para recomendação:

- instituição de vínculo: consultor e coordenador de projeto não podem atuar profissionalmente no mesmo departamento da instituição, nem na mesma instituição em uma mesma cidade;
- membros de equipe de projeto: o consultor não pode avaliar a proposta na qual conste como membro da equipe do projeto;
- níveis do consultor e do coordenador do projeto: consultores não podem possuir níveis inferiores ao do coordenador do projeto; e
- membro de comitê de assessoramento: o consultor não pode avaliar propostas vinculadas ao comitê do qual ele seja membro titular com mandato corrente.

Módulos do sistema de recomendação:

A figura 4.2 exibe os principais módulos do sistema de recomendação em uso no CNPq.

- carga do banco de consultores: atualiza o banco de consultores incluindo e excluindo bolsistas de produtividade conforme suas bolsas sejam implementadas ou encerradas, além disso, realiza atualizações decorrentes de alterações nos currículos dos consultores cadastrados no banco de consultores;
- configuração dos pesos dos parâmetros: registra os pesos dos parâmetros que serão usados na recomendação;

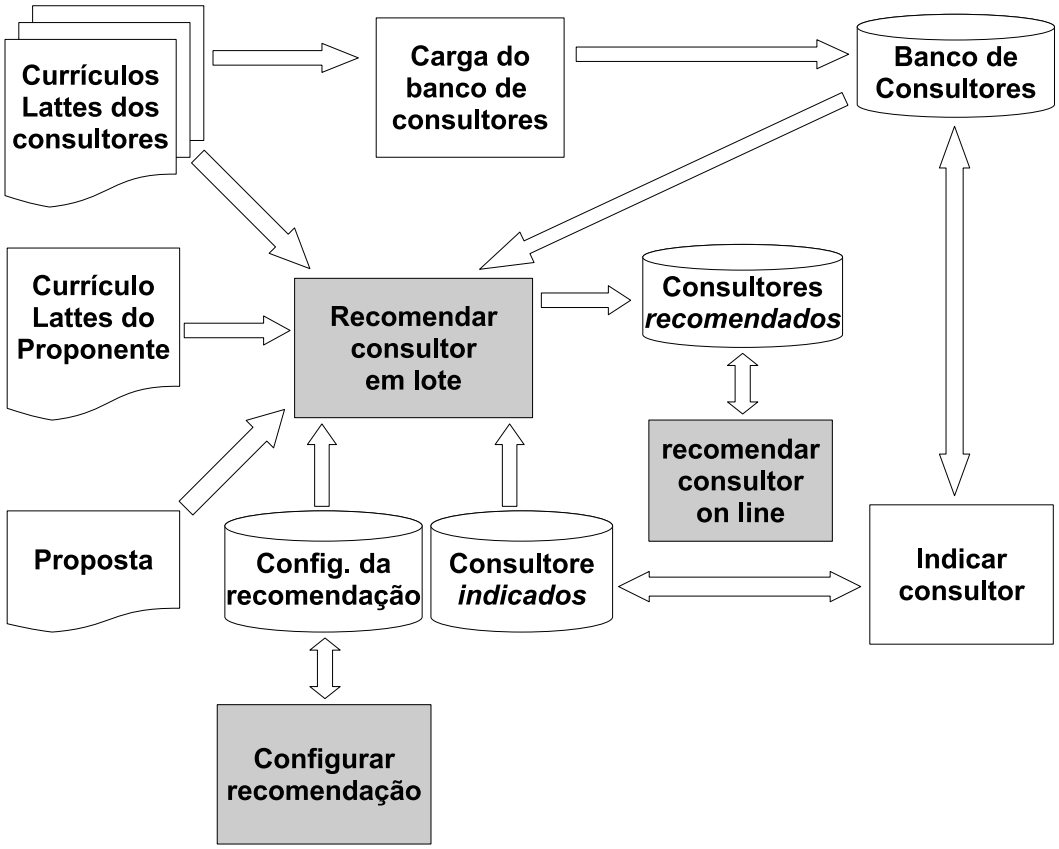

Figura 4.2: Módulos do sistema de recomendação

- recomendação em lote: realiza a recomendação em lote para todos os projetos submetidos em um Edital/Chamada;
- recomendações *on line*: descarta as recomendações geradas e recomenda novos consultores para um projeto específico por solicitação dos técnicos.

4.1.1 Vantagens

As principais vantagens do sistema de recomendação em uso no CNPq são:

- o sistema sugere consultores que eventualmente não seriam lembrados em indicações sem recomendação;
- mais agilidade na seleção de consultores; e
- melhor distribuição da carga de trabalho entre os consultores indicados.

4.1.2 Dificuldades e limitações

Algumas limitações da abordagem atual foram levantadas, a partir de sua análise e, também, de contatos com técnicos usuários do sistema:

- os critérios de similaridade entre os currículos dos consultores e dos proponentes não levam em consideração atributos semânticos,
- os critérios de adequação dos consultores para avaliação das propostas não levam em consideração atributos semânticos,
- os pesos para cálculo dos índices de similaridade são determinados empiricamente, sendo difícil mensurar os impactos de alterações nos mesmos sobre os resultados,
- não há suporte para perguntas do usuário como “Por que o consultor X foi recomendado?” ou “Por que o consultor Y não foi recomendado?”,
- a inclusão de novos critérios de similaridade é complicada,
- o tempo de resposta é alto,
- vinculação forte com o modelo de dados transacional, assim qualquer alteração no modelo de dados tem impacto elevado no mecanismo de sugestão de consultores;
- proponentes e os consultores com produção em coautoria não detectada por problemas de grafia na citação bibliográfica;
- a área de atuação dos consultores mais adequados às vezes é diferente da área da proposta;
- consultores e proponentes podem estar relacionados por orientação não detectada por problema de grafia nos nomes declarados;
- dificuldades para usar um critério excludente como um critério ponderado e vice-versa;

- dificuldades para inclusão de novos critérios de similaridades de consultores;
- a avaliação dos currículos dos envolvidos é realizada apenas com base na coincidência de palavras provida pelo SGDB sem tratamento para os fenômenos linguísticos;
- o sistema não possui capacidade de aprendizagem, e não recebe retroalimentação, não considera os índices de aceitação ou rejeição das sugestões já realizadas e não utiliza o histórico de indicações já realizadas pelos técnicos;
- não leva em conta as solicitações de dispensa de emissão de parecer, nem as justificativas apresentadas pelos consultores indicados;
- não considera a participação dos consultores e dos proponentes em de grupos de pesquisa.

4.1.3 Avaliação do sistema de recomendação em uso no CNPq

O mecanismo atual para recomendação automática de consultores *ad-hoc* foi avaliado em função de sua efetiva utilização pelos técnicos do CNPq e pela aceitação e emissão de parecer pelos consultores indicados. A Figura 4.3 resume as indicações realizadas pelos técnicos do CNPq, com base no aceite de recomendações obtidas com o mecanismo atual, no período de setembro de 2006 a 2009. Foram avaliadas 106.501 propostas com a participação de consultores *ad-hoc*, sendo que o mecanismo atual gerou 1.443.114 recomendações automáticas de consultores. Para avaliar as propostas, os técnicos do CNPq realizaram 231.528 indicações de consultores *ad-hoc*, média de 2,17 consultores por proposta. Dentre os consultores indicados, 156.219 (67,47%) foram previamente recomendados pelo mecanismo atual de recomendação de ad-hoc. Dos consultores recomendados e indicados até dezembro de 2009, foi enviado convite para emissão do parecer para 155.571 (67,19% do total), dos quais 148.949 aceitaram avaliar a proposta (64,33% do total). O número

de consultores recomendados e indicados que emitiram o parecer foi de 126.531 (54,65% do total), até dezembro de 2009.

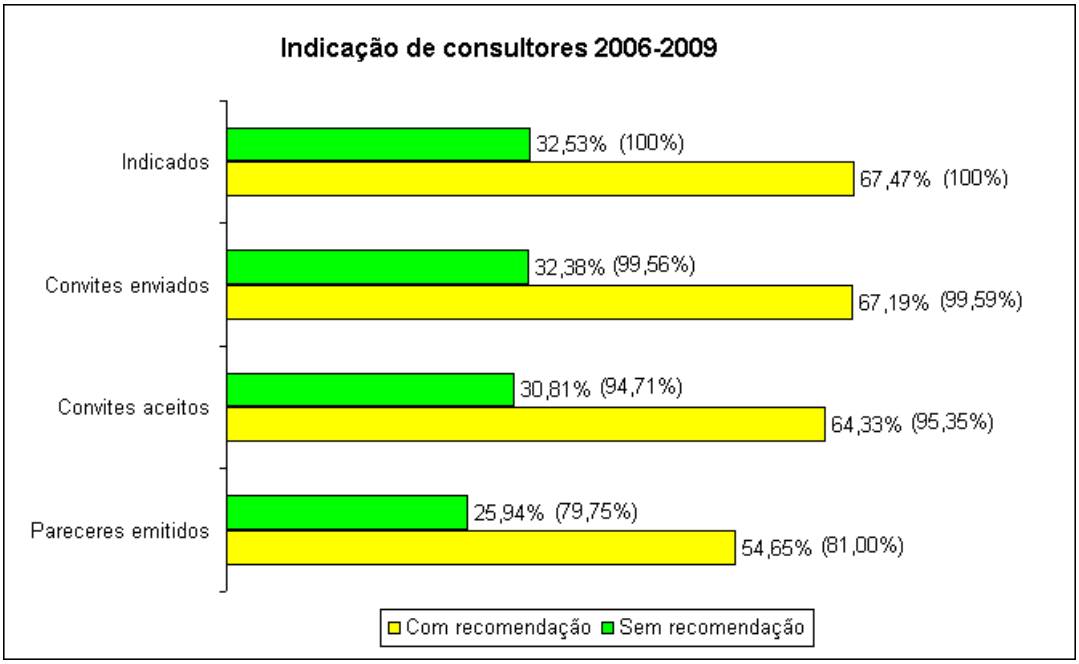

Figura 4.3: Estatística de consultores indicados

O desempenho do mecanismo atual, do ponto de vista da aceitação da recomendação pelos técnicos do CNPq, pode ser considerado como 67,47%, pois essa é a percentagem das indicações de consultores *ad-hoc* com origem em recomendações do sistema atual. Se for levado em conta que o objetivo final é que o consultor selecionado avalie a proposta, essa percentagem cai para 54,65%. No entanto, o desempenho relativo efetivo do sistema tual pode ser considerado 81,00% calculado como a razão entre 54,65 e 67,47%. De forma análoga, o desempenho das indicações feitas apenas pela equipe técnica é de 79,75%, ou seja, a razão percentual entre 25,94% e 32,53%. Portanto, o sistema atual apresenta desempenho similar ao da equipe técnica do CNPq, na indicação de consultores *ad-hocs*.

A tabela 4.1 e a figura 4.4 apresentam o desempenho do sistema de recomendação atual. Esse sistema começou a ser utilizado em setembro de 2006 para o Edital Universal e passou a ser utilizado na análise de todos os editais pelos quais o CNPq é responsável pela seleção, a partir de 2007, inclusive. Na tabela 4.1 pode-se

observar um crescimento do número de consultores recomendados e de consultores indicados (com ou sem recomendação do sistema atual), com desempenho efetivo máximo de 84,30% do sistema atual em 2007.

Ano	**Consultores recomendados**	**Consultores indicados (C.I.)**	**% dos C.I. recomendados (C.R.I.)**	**C.R.I. com convites enviados (%)**	**C.R.I. que aceitaram o convite (%)**	**C.R.I. que emitiram o parecer (%)**	
						Final	**Efetivo**
2006	82.375	16.321	69,68	69,58	66,52	51,02	73,22
2007	279.381	59.338	73,57	73,52	71,12	62,02	84,30
2008	590.591	78.824	68,18	67,61	64,32	56,16	82,58
2009	450.264	74.164	61,59	61,21	58,46	48,42	78,61
Média			**67,47**	**67,19**	**64,33**	**54,65**	**81,00**

Tabela 4.1: Desempenho anual da abordagem atual de recomendação

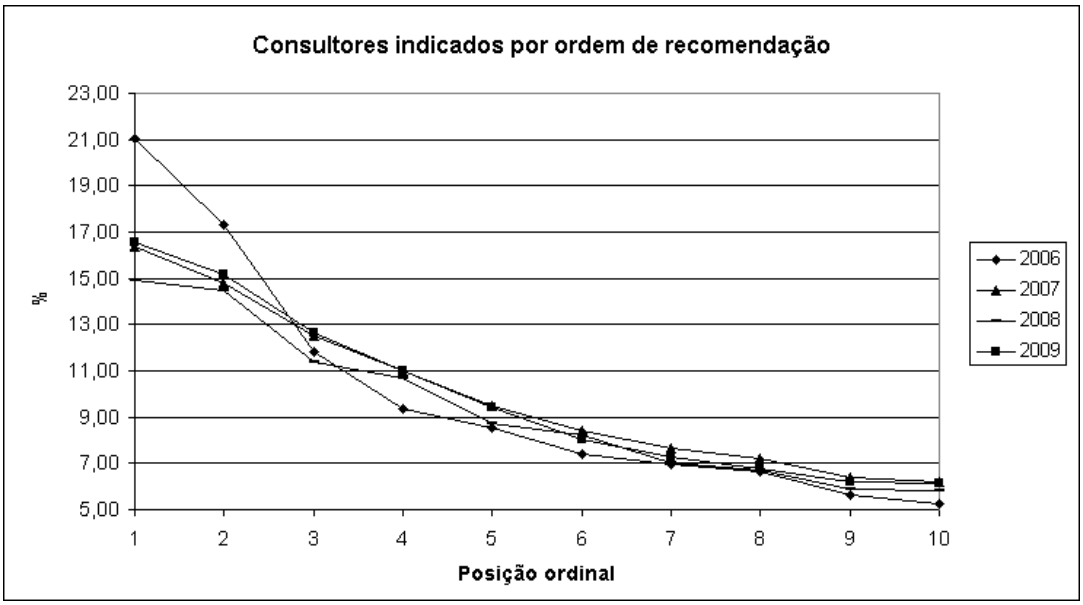

Figura 4.4: Consultores indicados por ordem de recomendação

As figuras 4.4 e 4.5 apresentam os percentuais de consultores indicados pelos técnicos do CNPq a partir de recomendação do sistema e de consultores recomendados pelo sistema, indicados e que emitiram pareceres para as propostas, respectivamente. Essa distribuição percentual está organizada de acordo com a posição ordinal da recomendação apresentada aos técnicos do CNPq, respectiva-

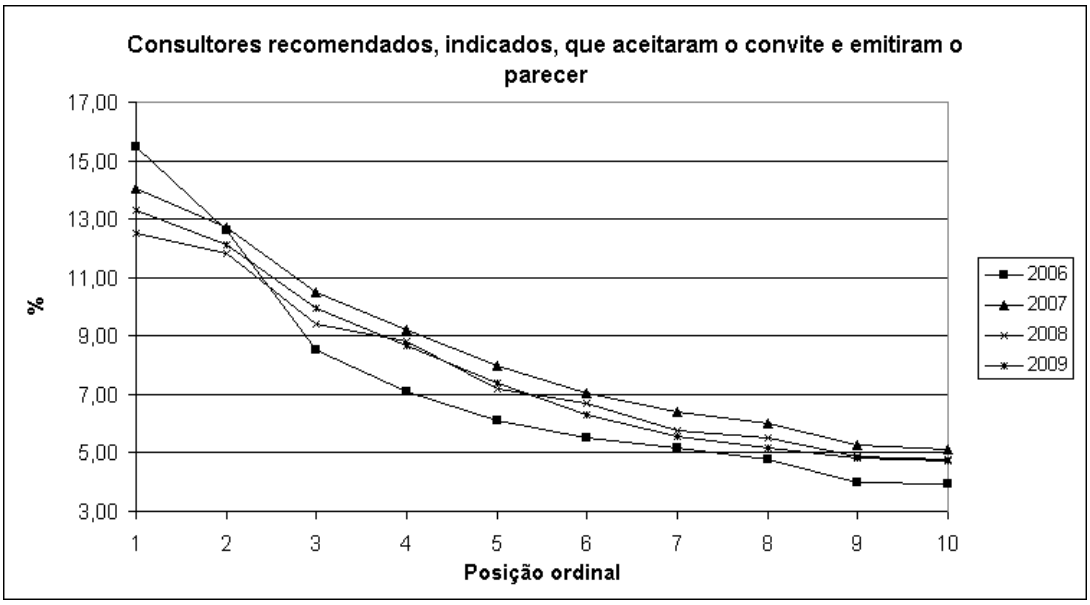

Figura 4.5: Consultores que emitiram o parecer por ordem de recomendação

mente. Constatou-se que as recomendações convertidas em indicações e pareceres tem a ser aquelas com maiores índices de similaridade (figura 4.4). No período considerado, a taxa de aceitação das recomendações feitas pelo sistema pela equipe técnica do CNPq pode ser considerada 67,47% pois esse é o percentual dos consultores indicados como *ad-hoc* que foram selecionados a partir de recomendações do sistema atual. Como apenas 54,65% chegaram ao fim do processo e emitiram pareceres, o desempenho efetivo médio do sistema atual pode ser considerado 81%. Assim, embora a aceitação do sistema é relativamente baixa, a aceitação de suas recomendações, em 81% dos casos, são bem sucedidas.

Capítulo 5

Metodologia proposta

Este capítulo apresenta a metodologia proposta para recomendação automática de consultores para avaliação de propostas submetidas ao CNPq, considerando as necessidades de se adequar ao sistema em uso no CNppq. A metodologia proposta para recomendação de consultores *ad-hoc* para avaliar projetos submetidos ao CNPq utiliza técnicas de mineração de textos e VSM para construir perfis dos pesquisadores e das propostas. Os perfis são relacionados via matrizes de similaridade. A recomendação de consultor *ad-hoc* é precedida de uma análise de conflitos de interesses.

5.1 Foco de atenção

Este projeto se concentra na recomendação de consultores avaliadores de propostas, submetidas a um processo de avaliação, cujo objetivo final é sua implementação. Essas propostas podem ser de naturezas diversas como: publicação de um artigo, livro ou capítulo de livro para publicação; obtenção de recursos de agências de fomento, públicas ou privadas, para execução de projeto, realização ou participação de evento; concessão de bolsa de estudos ou de apoio a pesquisa e assim por diante

Os pressupostos assumidos nesta proposta são que:

- a demanda por recursos através das propostas é superior à oferta, tornando necessário uma seleção de propostas mais adequadas ou mais viáveis;
- os responsáveis pela seleção das propostas a serem implementadas não detêm todo o conhecimento necessário para avaliar as propostas, fazendo com que seja necessário recorrer a especialistas que atuemo como consultores avaliadores;
- os consultores compõem um conjunto conhecido de pessoas habilitadas para tanto;
- os consultores podem rejeitar a indicação para avaliar uma proposta específica por considerarem-se impedidos, incapazes ou impossibilitados para avaliar a proposta de projeto;
- os pareceres dos consultores avaliadores são de mérito e serão utilizadas subsidiariamente em etapas posteriores de avaliação, onde outros critérios não técnicos podem ser aplicados como, por exemplo, prioridades empresariais, tendências de mercado, políticas de investimento, políticas de governo;
- consultores e proponentes possuem Currículo Lattes; e
- as propostas estão expressas em língua portuguesa;
- palavras-chave são descritores informados pelos pesquisadores e devem, tanto quanto possível, ser usadas sem alterações que as descaracterize.

A recomendação de consultores para avaliação de propostas requer que algumas etapas sejam observadas para sua realização:

1. Definição dos atributos (descritores) relevantes a serem considerados:
    - na composição dos perfis dos consultores e proponentes, e
    - na composição dos perfis das propostas.

2. Definição da forma como os descritores serão combinados no cálculo da similaridade:

    - **critérios positivos** - sua ocorrência aumenta a probabilidade de recomendação de um consultor,
    - **critérios negativos** - sua ocorrência diminui a probabilidade de recomendação de um consultor, e
    - **critérios excludentes** - sua ocorrência impede a recomendação de um consultor.

3. Escolha dos critérios de similaridade:

    - entre os currículos dos consultores e dos proponentes, e
    - entre currículos dos avaliadores e as propostas

4. Elaboração dos procedimentos de carga:

    - dos perfis dos consultores,
    - dos perfis dos proponentes, e
    - dos perfis das propostas.

5. Identificação de critérios intervenientes a serem ponderados na recomendação de um consultor:

    - carga de trabalho atribuída aos consultores,
    - número de recomendações por consultores, e
    - limite de corte para os critérios de similaridades:

        entre atributos,

        entre consultores e propostas, e

        entre consultores e proponentes.

6. Critérios de sucesso das recomendações:

- recomendações acatadas pelos técnicos,
- solicitações de dispensa de emissão de parecer pelos consultores, e
- emissão de parecer pelos consultores indicados.

5.2 Detalhamento da Solução Proposta

As características da metodologia de recomendação de consultores *ad-hoc* proposta neste projeto são:

1. Filtragem baseada em conteúdo:
   - perfis dos consultores,
   - peris dos proponente, e
   - perfis das propostas
2. Formas de combinar os critérios de recomendação:
   - **positivos** – sua ocorrência aumenta a probabilidade de recomendação de um consultor
   - **negativos** – sua ocorrência reduz a probabilidade de recomendação de um consultor
   - **excludentes** – sua ocorrência impede a recomendação de um consultor
3. representação dos perfis por meio do modelo de espaço vetorial VSM
   - construir a base do espaço vetorial por área do conhecimento com os termos extraídos dos currículos dos consultores e desprezar os termos presentes dos currículos dos proponentes ou nas propostas que não constem no currículo dos consultores
   - construir dois modelos VSM: um para pesquisadores (consultores e proponentes) e outro para propostas

- normalizar os termos extraídos dos dados textuais usados na construção do VSM, exceto no caso de palavras-chave usadas como descritores as quais devem sofrer alterações mínimas
- utilizar TF-IDF para cálculo dos pesos dos termos
- construir os vetores por a área do conhecimento de acordo com as informações dos currículos e das propostas
- converter os vetores para norma unitária
- construir matrizes de similaridades entre consultor e proponente e entre consultor e proposta

4. conflitos de interesses

    - produção científica e tecnológica conjunta
    - relacionamento orientador-orientando entre consultor e proponente
    - consultor membro da proposta de projeto
    - consultor concorrendo com projeto no mesmo conjunto de projetos a ser a avaliado
    - consultor membro do Comitê Consultivo que avalia a proposta
    - consultor e proponente atuam na mesma instituição
    - consultor e proponente membros do mesmo grupo de pesquisa
    - consultor vinculado à instituição de execução da proposta

A Figura 5.1 exibe um diagrama de blocos com os principais módulos requeridos para o sistema de recomendação automática de consultor avaliador de proposta.

O módulo de pré-processamento é responsável por: (i) extração e tratamento inicial dos dados que serão utilizados como descritores estruturados, textuais e semi-estruturados; (ii) realizar de conversões de formato, padronizações, substituição de termos por sinônimos – se houver um dicionário disponível; (iii) remoção de *stop*

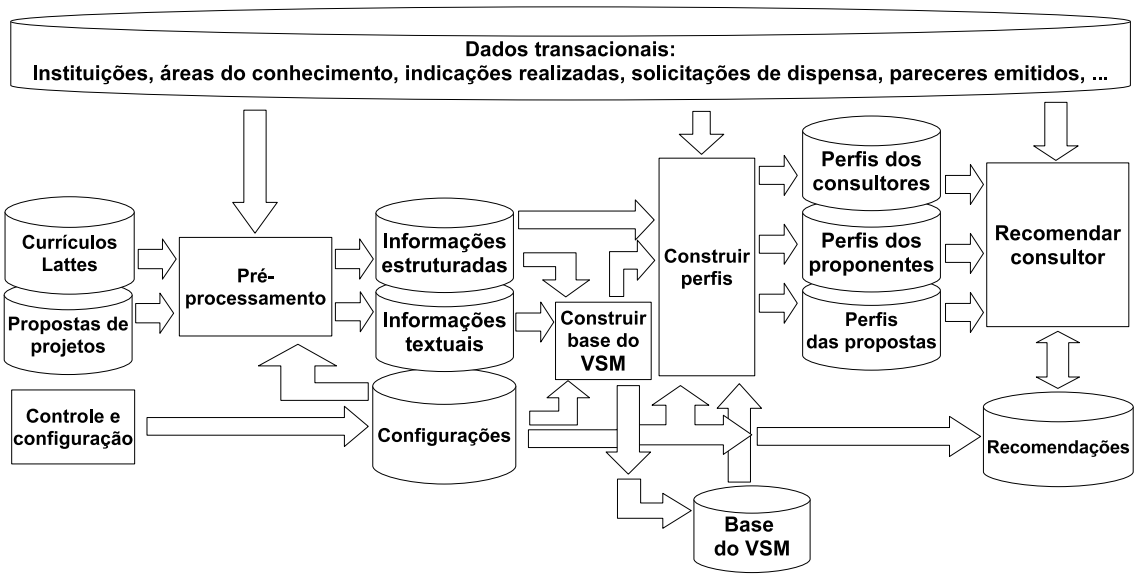

Figura 5.1: Módulos principais da recomendação de consultor *ad-hoc* proposta

words; (iv) lematização de termos e (v) construção de estruturas intermediárias necessárias.

O módulo de construção do VSM é responsável por: (i) construir a base do VSM com termos extraídos dos currículos dos consultores, (ii) aplicar as regras de redução da dimensão da base do VSM, (iii) aplicar as regras de atribuição de pesos aos termos do VSM, (iv) normalizar os vetores resultantes e (v) construir as matrizes de similaridades consultor-consultor, consultor-proponente, consultor-proposta.

O módulo de construção dos perfis é responsável por combinar as informações estruturadas e os vetores VSM em uma representação da constituição dos perfis, para fins de cálculo das similaridades.

O módulo de recomendação de consultor é responsável por: (i) recuperar e recomendar os N consultores com maiores índices de similaridade em relação à proposta e ao proponente; (ii) manter um histórico das recomendações realizadas e das ações dos técnicos em aceitar ou rejeitar as recomendações; (iii) manter um histórico das ações dos consultores ao rejeitar o convite para emissão de parecer, ou emitir o parecer; (iv) responder às perguntas do usuário relativas ao motivo da recomendação, ou não, de um consultor para avaliar uma proposta.

Critérios de sucesso das recomendações realizadas

Os critérios de sucesso podem ser estruturados em níveis de acordo com as ações dos técnicos ou dos consultores. O primeiro nível de sucesso, é a indicação do consultor pelo técnico, entretanto se a essa indicação não for avaliada de forma adequada pelo técnico, poderá ser rejeitada pelo consultor. Dependendo dos motivos para rejeição da indicação, o pedido de dispensa pode não ser acatado. Por isso o segundo nível de sucesso deve ser avaliado em função da emissão do parecer pelo consultor indicado.

5.3 Detalhamento da abordagem proposta

O primeiro passo consiste em selecionar os critérios para determinação similaridade entre consultores e propostas. Os critérios a serem utilizados devem se enquadrar em um dos três grupos já apresentados anterior: positivos, negativos ou excludentes.

Formalmente, a recomendação de consultores para avaliação de propostas pode ser vista com uma função de que associa um índice de similaridade ($Score$) a um par ordenado composto por um consultor e uma proposta. A recomendação consiste em escolher os pares com maiores índices associados.

$$Score(C_i, P_j) = Neg(C_i, P_j)Sim(C_i, P_j)$$

C_i é um currículo de um consultor, P_j é uma proposta. $Sim(C_i, P_j)$ é uma função de similaridade entre o consultor C_i e uma proposta P_j. $Neg(C_i, P_j)$ é uma função cujos valores de retorno são zero ou um. Zero indica conflito de interesses entre o consultor C_i e alguma característica da proposta P_j e deve implicar na não recomendação do consultor.

A similaridade $Sim(C_i, P_j)$ entre um consultor e uma proposta pode ser decomposta em dois índices combinados por algum critério, ou função F:

- ($SimC(C_i, C_j)$), similaridade entre os perfis curriculares dos pesquisadores i e j; e
- ($SimP(C_i, P_j)$), similaridade entre o perfil curricular do consultor i e a proposta j.

$$Sim(C_i, P_i) = F(SimC(C_i, C_j), SimP(C_i, P_j))$$

Os atributos (descritores) a serem utilizados nos critérios de similaridade podem ser estruturados ou textuais. Para os atributos estruturados deve existir uma função de comparação que permita calcular um índice de semelhança entre eles de tal forma que: para o critérios positivos, o índice aumente com a semelhança; para os critérios negativos, o índice diminua com o aumento da semelhança; e para os critérios excludentes, o índice seja zero sempre que algum conflito de interesses for detectado ou um, caso contrário.

Os atributos textuais merecem uma atenção especial, pois guardam relações semânticas difíceis de serem analisadas por métodos computacionais ou estatísticos, além de implicarem em muito espaço de armazenamento e tempo de processamento. Devem ser escolhidos atributos que sejam relevantes para a construção dos perfis como palavras-chaves, resumos, títulos etc.

Como os consultores correspondem a um conjunto de referência para a recomendações, o primeiro passo para extração dos descritores textuais é estabelecer uma base de termos a partir dos atributos textuais extraídos dos currículos dos consultores. Com isso, a dimensão da base de termos não crescerá com a adição de novas propostas e novos currículos de proponentes. Essa base pode ser reconstruída periodicamente em função das atualizações dos currículos dos proponentes. Uma alteração incremental da base em função da inclusão, exclusão e alterações nos currículos dos consultores também é possível, embora mais complicada, pois têm

impactos nas representações vetoriais dos currículos dos pesquisadores e das propostas. A solução desse problema está fora do escopo deste trabalho.

Uma vez construída uma base de termos, extraídos dos currículos dos consultores, são construídos vetores VSM para representação dos currículos dos consultores, dos currículos dos proponentes e das propostas nessa mesma base, evitando, assim, a representação desnecessária de termos nos VSM das propostas e dos proponentes que não constam nos VSM dos consultores, não contribuem para identificar semelhanças com os consultores. Exceto no caso de uso de um dicionário, tesauro ou ontologia.

Dessa forma, os currículos e as propostas podem representados como uma tupla de descritores ($Desc$):

$$C_i = (Desc_{C_i,1}, Desc_{C_i,2}, \ldots, Desc_{C_i,k}, \vec{v}_{C_i})$$
$$P_j = (Desc_{P_j,1}, Desc_{P_j,2}, \ldots, Desc_{P_j,n}, \vec{v}_{P_j})$$

onde $Desc_{C_i,p}$ é o p-ésimo descritor estruturado do currículo i e $\vec{v}_{C_i}$ é a sua representação no VSM, $Desc_{P_j,k}$ é o k-ésimo descritor estruturado da proposta j e $\vec{v}_{P_j}$ sua representação no VSM.

Visando reduzir o tempo de resposta, podem ser construídas as matrizes de similaridades baseadas nos VSM para os pares consultor-consultor, consultor-proponente e consultor-proposta. As matrizes de similaridade consultor-proposta e consultor-proponente serão usadas para o cálculo da similaridade ente consultores e propostas. A matriz de similaridades consultor-consultor é útil para encontrar outros consultores candidatos para serem recomendados. A ideia é que, consultores com perfis semelhantes podem realizar tarefas semelhantes.

As similaridades entre os perfis dos currículos dos consultores e dos proponente e, entre currículos dos consultores e as propostas, são obtidas mediante a combinação das similaridades entre os atributos comparáveis, ponderados por um peso arbitrário. Dois atributos são comparáveis se pertencerem ao mesmo domínio

semântico como por exemplo área do conhecimento, conjunto dos pesquisadores, conjunto das instituições, Conjunto dos Comitês de Assessoramento, localização, etc.

A utilização da distribuição da carga de trabalho entre os consultores como critério de recomendação depende de um parâmetro que pode ser obtido somente no momento da indicação, pois depende de dados dinâmicos. Para calcular o *score* da carga de trabalho dos consultores em relação a um conjunto de propostas, é necessário determinar o número médio ($\overline{n}$) de propostas a serem avaliadas pelos consultores disponíveis para recomendação no início do processo e, a cada recomendação, o número (n_i) de propostas que foram distribuídas para cada consultor candidato a ser recomendado. Esse *score* deve ser tal que, consultores com menor carga de trabalho em relação à media tenham peso maior na seleção. Para esse critério ainda pode ser determinado um limite de corte, de forma que um consultor não receba mais do que um determinado número de propostas do conjunto de propostas a ser avaliado.

A título de exemplo, considere que p seja o peso atribuído à carga de trabalho do consultor. Seu *score* S_t pode ser calculado por

$$S_t(n_i) = (\frac{\overline{n} - n_i}{\overline{n}})p$$

S_t foi escolhido como uma função linear sobre n_i. Outras construções são possíveis, dependendo da forma como a similaridade $Sim(C_i, P_j)$ será calculada.

Essa formulação é útil por que é uma função não crescente, isto é, se $n_i \leq n_j$, então $S_t(n_i) \leq S_t(n_j)$, além disso, para $n_i = \overline{n}$, $S_t(n_i) = p$ e, se $n_i > \overline{n}$, então $S_t(n_i) < 0$. Isso faz com que o índice de similaridade final seja penalizado com valores negativos quando a carga de trabalho do consultor for superior à média.

$$SimC(C_i, C_j) = \sum_k Peso_k SimC_{Desc}(Desc_{C_i,k}, Desc_{P_j,k})$$
$$SimP_{C_i,P_j} = \sum_k Peso_k SimP_{Desc}(Desc_{C_i,k}, Desc_{P_j,k})$$

Onde $SimC_{Desc}$ e $SimP_{Desc}$ são funções atribuem um índice de similaridade a atributos k dos objetos que estão sendo comparados, desde que os atributos em comparação pertençam ao mesmo domínio semântico.

A figura 5.2 representa a estrutura lógica da metodologia implementada para fins de testes e validação. Alguns detalhes foram omitidos, como por exemplo: a utilização, ou não de representação XML; a necessidade de um lematizador; aplicação de filtros de *stop words*; uso de ferramentas de apoio como dicionários, tesauros e ontologias.

Os seguintes parâmetros foram utilizados nas simulações e testes realizados:

Critérios de similaridades positivos

- Proximidade da área do conhecimento, subárea e especialidade de atuação do consultor e da proposta.
- Proximidade da área do conhecimento, subárea e especialidade de atuação do consultor e do proponente.
- Comitê de Assessoramento de vínculo do consultor e de julgamento da proposta.
- Nível do consultor superior ao do proponente.
- Proximidade entre os vetores VSM de representação do consultor e da proposta.
- Proximidade entre os vetores VSM de representação do consultor e do proponente.

Critérios de similaridades negativos:

- Instituições de vínculo do consultor e do proponente, se em instituições ou em cidades diferentes.
- Instituições de vínculo do consultor e de execução da proposta, se em instituições ou em cidades diferentes.

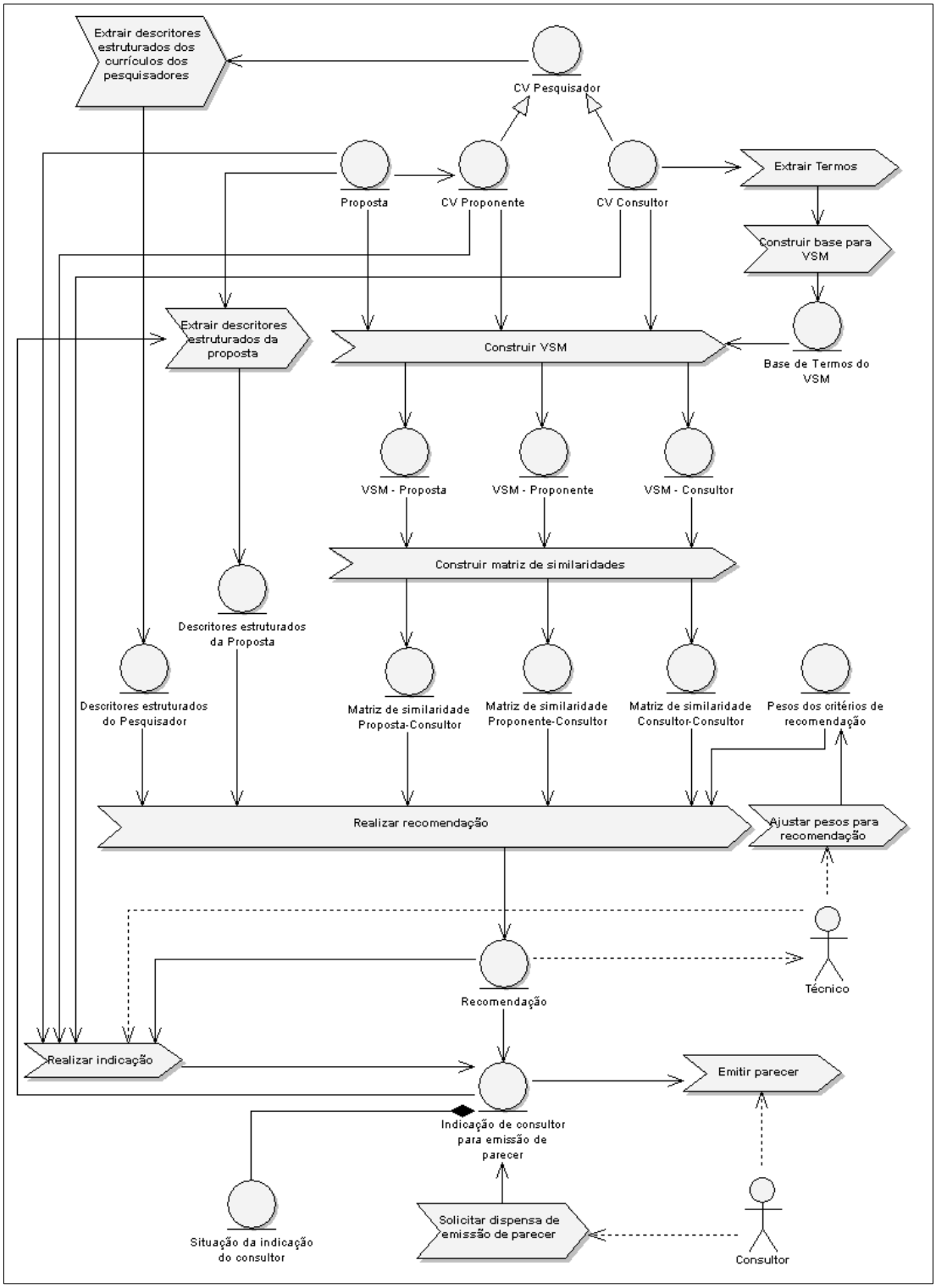

Figura 5.2: Diagrama de blocos

- Número de propostas para as quais o consultor já foi indicado em relação ao número médio de propostas por consultor dentro do edital/chamada.

Critérios excludentes:

- Mesmas instituições de vínculo do consultor e do proponente, se na mesma cidade ou departamento.
- Mesma Instituição de vínculo do consultor e de execução da proposta, se na mesma cidade ou departamento.
- Consultor membro da equipe de projeto.
- Nível do consultor inferior ao do proponente.
- Consultor membro do Comitê de Assessoramento que vai julgar a proposta.
- Coautoria em produção científicas entre consultor e proponente.
- Relacionamento orientador-orientando entre consultor e proponente e vice-versa.
- Consultor possui proposta concorrendo com a proposta a ser avaliada.

Capítulo 6

Resultados obtidos

Este capítulo apresenta a os resultados das simulações feitas com a metodologia proposta e os compara com o desempenho do sistema em uso no CNPq. Foram construídos três modelos de representação dos perfis usando o modelo de espaço vetorial VSM com o objetivo de identificar qual conjunto de dados é mais recomendável para uso na metodologia. Para redução do esforço computacional e da dimensão dos espaços vetoriais envolvidos, foram realizados testes paramétricos de descarte de termos de baixa frequência os currículos.

6.1 Construção dos perfis no modelo VSM

Foram realizados diversas simulações de construção dos VSM utilizando 12.451 currículos de consultores cadastrados no banco de consultores do CNPq e 39.901 propostas submetidas aos editais Universal MCT/CNPq de 2006, 2007 e 2008. Com esses montantes tornou-se evidente a existência de explosão de dimensionalidade do espaço vetorial para cálculo do modelo VSM. Para contornar esse problema, foram realizados os seguintes estudos paramétricos visando reduzir a dimensionalidade do espaço vetorial:

- consideradas áreas do conhecimento até o nível de especialidade. Essas áreas estão organizadas em quatro níveis (grande área, área, subárea, especialidade), na forma de uma tabela de áreas do conhecimento usadas por agências de fomento como o CNPq e a CAPES[1];
- aplicadas técnicas de pré-processamento de texto e estudos de determinação do número mínimo de ocorrências de atributos para serem considerados no modelo.

Para reduzir a dimensão do espaço vetorial no modelo VSM e avaliar a contribuição específica de cada atributo, foram construídos as seguintes representações VSM C1 a C3 para os currículos dos pesquisadores (consultores e proponentes) e VSM P1 a P3 para as propostas:

- *VSM-C1* -– Palavras-chave (**key**) da produção científica e tecnológica constantes dos currículos dos pesquisadores, nos últimos 5 anos. Esse espaço vetorial é representado na base C1, obtida com os atributos dos consultores.
- *VSM-C2* -– Termos extraídos das palavras-chave, título e especialidade da subárea da produção científica e tecnológica (**title**), nos últimos 5 anos. Esse espaço vetorial é representado na base C2, obtida com os atributos dos consultores.
- *VSM-C3* -– Termos extraídos do nome e especialidade da subárea da última titulação do pesquisador (**major**). Esse espaço vetorial é representado na base C3, obtida com os atributos dos consultores.
- *VSM-P1* -– Palavras-chave da proposta de projeto, representadas na base C1.
- *VSM-P2* -– Termos extraídos das palavras-chave, título, resumo e especialidade da subárea da proposta, representados na base C2.

[1]Tabela de áreas do conhecimento do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq. Disponível em http://www.cnpq.br/areasconhecimento/index.htm

- *VSM-P3* -– Termos extraídos das palavras-chave, título, resumo e especialidade da subárea da proposta, representados na base C3.

A base C1 é utilizada para representar o espaço vetorial obtido com as palavras-chave da produção dos pesquisadores, ou seja VSM-C1. A estrutura VSM-P1 consiste na representação, nesta base, das palavras-chave obtidas nas propostas de projetos. Essas estruturas são formadas por frequências ponderadas pela métrica TF-IDF, normalizadas para que o comprimento de cada vetor desse espaço seja unitário. A base C2 é utilizada para construção das representações vetoriais VSM-C2 (produção dos consultores e proponentes) e VSM-P2 (propostas de projetos). Essas estruturas vetoriais receberam tratamento similar às palavras-chave e também são convertidas para norma unitária. O mesmo raciocínio se aplica à base C3 utilizada para representações VSM-C3 (titulação dos pesquisadores) e VSM-P3 (propostas de projetos).

		Proponentes			**Propostas**			**Nome do modelo**
	VSM	**C1**	**C2**	**C3**	**P1**	**P2**	**P3**	**(soma de cossenos)**
Consultores	**C1**	C1C1			C1P1			M-key
	C2		C2C2			C2P2		M-title
	C3			C3C3			C3P3	M-major

Tabela 6.1: Matrizes de similaridade construídas

A partir das representações vetoriais normalizadas foram obtidas matrizes de similaridades (tabela 6.1) entre consultores e proponentes (C1C1, C2C2 e C3C3) e entre consultores e propostas (C1P1, C2P2 e C3P3), através do cálculo do cosseno entre vetores. Dado um proponente, calculou-se o cosseno entre os vetores representando consultores e o proponente, e entre os vetores dos consultores e da proposta daquele proponente conforme o modelo VSM já descrito, esses valores foram utilizados no cálculo final da similaridade entre os perfis. As matrizes de similaridade relativas à mesma base do espaço vetorial (mesma linha na tabela 6.1), foram somadas com aplicação dos pesos de ponderação $Peso_{VSM-A}$ e $Peso_{VSM-B}$

(tabela 6.2) gerando três modelos que usados para representação dos consultores, proponentes (VSM C1 a C3) e propostas (VSM P1 a P3).

Na construção dos VSM foram aplicados os seguintes critérios:

1. uso do índice TF-IDF para o cálculo dos pesos dos termos no VSM, com descarte dos termos com peso igual a zero, e
2. normalização das coordenadas para obter vetores de comprimento igual a unidade.

Critério			**Peso**	**Parâmetro**
Número máximo de sugestões de consultor por proposta				10
Número máximo de propostas por consultor				Sem limite
Nível CNPq adotado para pesquisador sem bolsa PQ				2
Similaridade entre consultores e propostas (baseada em atributos estruturados)				
	Proposta a ser julgada pelo mesmo Comitê do consultor sugerido ($Peso_{comite}$)		0,3	
	Proposta na mesma subárea de conhecimento do consultor sugerido, mas em especialidade da subárea distinta ($Peso_{subarea}$)		0,3	
	Proposta na mesma especialidade de conhecimento do consultor sugerido ($Peso_{espec.}$)		0,5	
	Instituições diferentes: consultor sugerido, proponente e execução do projeto ($Peso_{inst.}$)		0,3	
	Nível do consultor ($Peso_{nivel}$)			
		Nível SR	0,9	
		Nível 1A	0,9	
		Nível 1B	0,85	
		Nível 1C	0,75	
		Nível 1D	0,7	
		Nível 2	0,6	
Similaridades entre perfis (baseada no modelo VSM)				
	Consultor-proposta ($Peso_{VSMA}$)		1	
	Consultor-proposta ($Peso_{VSMB}$)		1	

Tabela 6.2: Pesos e parâmetros para cálculo da similaridade

O cálculo do índice de similaridade final é realizado apenas para os consultores que atuam na mesma área do conhecimento em que se insere a proposta a ser avaliada. Esse índice é calculado como a soma ponderada das similaridades entre consultores e propostas (baseada na atribuição de pesos aos atributos estrutura-

dos) e das similaridades dos perfis (obtidas através dos modelos VSM). A tabela 6.2 apresenta parâmetros e pesos de ponderação utilizados no cálculo final dessas similaridades. Esses pesos são os mesmos utilizados no sistema de recomendação atual do CNPq, os quais foram escolhidos empiricamente. Se o critério não for atendido, é atribuído valor zero ao peso em questão. A expressão da similaridade final é denominada $Score_{final}$.

O nível do pesquisador é o nível da bolsa de produtividade em pesquisa do CNPq que o pesquisador possuía no momento do cálculo dos índices de similaridades. No caso de não possuir bolsa de produtividade e pesquisa e constar no banco de consultores, foi utilizado o nível ali registrado. Aos proponentes que não possuíam bolsa de produtividade em pesquisa no CNPq e nem constavam do banco de consultores, foi atribuído o nível padrão incial 2.

Os Comitês Assessores aos quais o consultor está associado são os comitês que preenchem pelo menos um dos requisitos: a) julgou sua bolsa de produtividade em pesquisa; b) no qual tem mandato ativo como membro; c) julgador da sua proposta de projeto; d) o consultor escolheu como seu comitê padrão para avaliação de suas propostas; e) foi informado pelo técnico do CNPq, quando cadastrou o pesquisador no banco de consultores *ad-hoc* do CNPq. No caso do consultor ter mais de um comitê associado, todos são levados em consideração, no cálculo da similaridade S_{comite}, isto é, receberá peso 0,3 se o comitê julgador da proposta a ser avaliada for um dos comitês ao qual ele está associado.

Foram utilizadas três classes de critérios: similaridade positiva (aumenta a probabilidade de recomendação), similaridade negativa (diminui a probabilidade de recomendação) e excludentes (impedem a recomendação). Esses critérios foram agrupados conforme abaixo:

Critérios de similaridade positivos

- Subárea do conhecimento ($S_{subarea} \in \{0, 1\}$)

1 se a subárea de atuação do consultor é a mesma da proposta mas de especialidades da subárea diferentes.

- Especialidade do conhecimento ($S_{espec.} \in \{0, 1\}$)

  1 se a especialidade da subárea de conhecimento de atuação do consultor é a mesma da proposta.

- Comitê Assessor de julgamento da proposta ($S_{comite} \in \{0, 1\}$)

  1 se consultor é vinculado ao mesmo CA.

- Nível do consultor ($S_{nivel} \in \{0.6, 0.7, 0.75, 0.85, 0.9\}$)

  consultor é considerado apenas se seu nível no CNPq for maior do que o do proponente,

  0.6 se nível 2; 0.7 se nível 1D; 0.75 se nível 1C; 0.85 se nível 1B, 0.9 se nível 1A ou SR.

- Modelo VSM consultor e proposta ($S_{VSM-A} \in [0, 1]$)

  similaridade dos perfis do consultor e proponente é calculada com base nas matrizes de similaridades C1P1 (entre as palavras-chave da produção do consultor nos últimos cinco anos e as palavras-chave contidas na proposta a ser avaliada), C2P2 (entre as palavras-chave, títulos e especialidades da subárea de conhecimento da produção do consultor nos últimos cinco anos e as palavras-chave, título, resumo e especialidade da subárea de conhecimento da proposta) e C3P3 (entre os títulos e especialidade da subárea da última titulação do consultor e as palavras-chave, título, resumo e especialidade da subárea de conhecimento da proposta), apresentadas na tabela 6.1.

- Modelo VSM consultor e proponente ($S_{VSM-B} \in [0, 1]$)

  similaridade dos perfis do consultor e proponente é calculada com base nas matrizes de similaridades C1C1 (entre as palavras-chave da suas produções nos últimos cinco anos do consultor e do proponente), C2C2 (entre as

palavras-chave, títulos e especialidades da subárea de conhecimento da produção do consultor nos últimos cinco anos do consultor e do proponente) e C3C3 (entre nome e especialidade da subárea de conhecimento da última titulação do consultor e do proponente), apresentadas na tabela 6.1.

Critérios de similaridade negativos

- Proximidade entre as instituições do consultor e da execução da proposta ($S_{inst.} \in \{0, 1\}$)

  0 se as instituições localizam-se na mesma cidade ou se são as mesmas mas localizadas em cidades diferentes.

Critérios excludentes

- Consultor e proponente vinculados à mesma instituição na mesma cidade.
- Consultor vinculado à mesma instituição de execução da proposta na mesma cidade.
- Nível do consultor menor do que o nível do proponente.
- Consultor membro da equipe de projeto.
- Consultor e proponente são membros do mesmo grupo de pesquisa.
- Consultor com mandato corrente no mesmo CA que julga a proposta.
- Consultor e proponente possuem produção científica ou tecnológica em conjunto nos últimos 5 anos.
- O consultor é, ou foi, orientador ou orientando do proponente.
- O consultor possui proposta submetida no mesmo edital e chamada da proposta a ser avaliada.

A ocorrência de pelo menos um critério excludente faz com que o consultor não mais seja considerado como recomendável para avaliar a proposta em questão.

O escore final de recomendação de um consultor para uma dada proposta é o índice de similaridade final entre o consultor e a proposta. Os consultores recomendados por essa metodologia foram aqueles com maiores escores.

O cálculo do escore final é dado por:

$$Score_{final} = \sum_{c \in Criterio} Peso_C$$

$$Criterio = \{nivel, subarea, espec., comite, inst., VSM - A, VSM - B\}$$

Na construção do VMS os seguintes tratamentos foram aplicados:

- extração dos termos da base pelo segundo nível no hierarquia das áreas do conhecimento (grande área, área, subárea, especialidade), de forma que a área atue como tópico, ou assunto, assim um mesmo consultor pode constar em mais de uma área com vetores distintos;
- remoção de *stop words* a partir de uma lista contendo termos em inglês e português, exceto para de palavras-chave;
- normalização dos termos:
    1. remoção de caracteres especiais,
    2. substituição de caracteres acentuados por não acentuados,
    3. substituição de caracteres com til e trema pelos mesmos sem os sinais gráficos,
    4. substituição de “ç” por “c”,
    5. remoção de excesso brancos,
    6. conversão para letras maiúsculas,
    7. exceto para palavras-chave tratadas como descritores, remoção dos sufixos *-NOS-EMOS*, *-SE-LHES*, *-LOS-EIS*, *-LHES-AS*, *-VOS-EIS*, *IME-*

TRIA, *-LHE-AS*, *-LAS-EI*, *-LHE-EI*, *-LHE-IA*, *-LHE-AO*, *-LOS-EI*, *-LO-EIS*, *-VOS-AO*, *-VOS-EI*, *-SE-LHE*, *IZACAO*, *-TE-EI*, *-TE-IA*, *-TE-AS*, *-LA-EI*, *-LA-AS*, *-LA-IA*, *-LO-AO*, *-LO-AS*, *-LO-EI*, *-LO-IA*, *-ME-AO*, *-ME-AS*, *-SE-AO*, *-SE-IA*, *-VOS-A* e *-LHE-A*, *CACAO*, *LOGIA*, *WINGS*, *ATION*, *-LHES*, *-SE-A*, *-LA-A*, *-LO-A*, *-ME-A*, *INGS*, *WING*, *-LHE*, *-LHA*, *-LHO*, *-VOS*, *-MOS*, *-NOS*, *-LOS*, *-LAS*, *-TE*, *-OS*, *-AS*, *-SE*, *-LO*, *-LA*, *-ME*, *-MO*, *-MA*, *-NA*, *-NO*, *-SE*, *-O*, *-A*,

8. exceto para palavras-chave tratadas como descritores, substituição dos prefixos: *ZATION* por *ZE*, *CATION* por *CA*, *AMENTE* por *A*, *TORES* por *OR*, *TORAS* por *OR*, *TIALS* por *TIAL*, *CALLY* por *C*, *ARES* por *AR*, *ISMS* por *ISM*, *TERS* por *TER*, *ADAS* por *ADO*, *ADOS* por *ADO*, *ANAS* por *ANO*, *ANOS* por *ANO*, *THMS* por *THM*, *ENTS* por *ENT*, *ESTS* por *EST*, *OUPS* por *OUP*, *PUTS* por *PUT*, *AGEM* por *A*, *EIRA* por *EIRO*, *ICAL* por *IC*, *IAS* por *IO*, *COES* por *CAO*, *AIS* por *AL*, *ICS* por *IC*, *RES* por *R*, *RAS* por *R*, *CAS* por *CO*, *COS* por *CO*, *NAS* por *NO*, *NOS* por *NO*, *ADA* por *ADO*, *ANA* por *ANO*, *ADA* por *ADO*, *ERS* por *ER*, *ALS* por *AL*, *ZED* por *ZE*, *TED* por *TE*, *ORS* por *OR*, *ADO* por *A*, *EMS* por *EM*, *ETS* por *ET*, *EMS* por *EM*, *CA* por *CO*, *AS* por *A*, *ES* por *E*, *IS* por *I*, *OS* por *O*, *US* por *U*, *NS* por *N*, *RR* por *R*, *MM* por *M*, *NN* por *N*, *EE* por *E*, *SS* por *S*, *OO* por *O*, *FF* por *F*, *LL* por *L*,

- descarte de termos com frequência igual a um, exceto para o VSM construído para a última titulação,
- uso do índice TF-IDF para o cálculo dos pesos dos termos no VMS, e
- normalização dos vetores do VSM.

Para verificação de coautoria, foi considerado que, dois pesquisadores são coautores se houver citação recíproca entre eles. Essa citação pode ser identificada de forma exata, por meio de chaves referenciadas ou pela utilização do nome completo.

Além disso, foi usada comparação por aproximação, usando a distância de Levenshtein para identificar as citações cruzadas aplicadas à citação propriamente dita. A distância de Levenshtein, ou distância de edição, é dada pelo número mínimo de inclusões, exclusões e substituições de caracteres necessárias para que um texto seja transformado em outro. Esse número foi convertido em um índice de similaridade, dividindo a distância de Levenshtein obtida pelo comprimento do maior texto e depois subtraindo de um. Esse índice é igual a um para textos iguais e igual zero, se todos os caracteres de um texto for substituído para igualar ambos [Poncelet et al., 2008].

6.1.1 Dados utilizados

Durante os testes forma utilizados dados do Edital Universal MCT/CNPq dos anos de 2006, 2007 e 2008, por abrangerem diversas áreas do conhecimento em cada edital. Os editais foram respectivamente, 02/2006, 15/2007 e 14/2008. A produção científica e tecnológica considerada para fins de extração dos dados textuais foram as dos últimos cinco anos contados retroativamente a partir do ano do edital. Dessa forma foram desprezadas as informações mais recentes que não estariam disponíveis na ocasião em que a proposta foi encaminhada para análise pelos consultores. Os dados utilizados durante a fase de experimentação resume-se em.

- 12.451 consultores
- ano 2006 – 12.233 propostas
- ano 2007

  até R$ 20.000,00 – 6.236 propostas

  de R$ 20.001,00 até R$ 50.000,00 – 6.803 propostas

  de R$ 50.001,00 até R$ 150.000,00 – 2.985 propostas

- ano 2008

  até R$ 20.000,00 – 4.623 propostas

  de R$ 20.001,00 até R$ 50.000,00 – 4.572 propostas

  de R$ 50.001,00 até R$ 150.000,00 – 2.449 propostas

Os testes finais foram realizados com as propostas enquadradas na primeira faixa do Edital Universal 142008 (ano 2008), para utilização de dados curriculares mais recentes.

Para os valores dos pesos aplicados no cálculo das similaridades, foram utilizados os mesmos valores correspondentes em uso no sistema de recomendação atual (tabela 6.2). Procurou-se utilizar também o mesmo conjunto de descritores em uso atualmente no CNPq.

6.2 Avaliação dos resultados

Verificou-se redução significativa no tamanho das bases dos VSM quando foram descartados termos de baixa frequência nos currículos dos consultores. O impacto na dimensão da base varia conforme os atributos escolhidos para construção do VMS. A figura 6.1 mostra o crescimento do percentual de currículos não recuperados pelo modelo VSM testado em função do número determos de baixa frequência descartados. Pode-se observar que as palavras-chave (modelo M-key) produzem um VSM maior do que a produção científica (M-major) e que a última titulação (modelo M-major) produz o espaço vetorial menor. Isso acontece por que as palavras-chave foram tomadas como descritores, produzindo uma combinação maior de ocorrências, enquanto para a formação e última titulação foi utilizada técnica de normalização de termos para redução da dimensão do espaço vetorial.

Após a aplicação das técnicas de pré-processamento de texto descritas, as cardinalidades das bases obtidas para os modelos M-key, M-title e M-major foram,

Frequência mínima de termos extraídos do currículo para descarte do termo	Dimensão da base		
	M-key	**M-title**	**M-major**
0	267.259	225.206	22.920
1	88.025	154.166	811
2	47.226	98.146	54
3	30.631	66.668	10
4	21.849	52.221	6
5	16.576	46.736	0
6	12.996	36.267	0
7	10.468	33.235	0
8	8.641	29.396	0
9	7.220	26.188	0

Tabela 6.3: Redução de dimensional dos VSM x frequência de descarte de termos

respectivamente: 225.206, 267.259 e 22.920 (tabela 6.3). A construção de matrizes de similaridades com essas dimensões requer um esforço computacional (em termos de processamento, armazenamento e tempo de resposta) que foi considerado proibitivo e inviável com os recursos disponíveis. Face a esses fatos, foi estudada a sensibilidade da abordagem proposta à redução da cardinalidade das bases VSM por descarte de termos. O impacto do descarte de termos na redução da dimensão das bases variou conforme os atributos escolhidos para construir os modelos VSM M-key, M-title e M-major (Tabela 1). Por exemplo, com o descarte de termos de frequência unitária, essas dimensões passaram para 88.025, 154.166 e 811. Com o descarte de termos com frequência menor ou igual a 5, essas dimensões foram drasticamente reduzidas para 10.468, 36.267 e 0. Portanto, não mais era possível construir modelos VSM com informação sobre a titulação do pesquisador.

A quantidade de currículos não recuperados cresceu com o aumento do número de termos descartados (figura 6.1). O descarte de termos da titulação mostrou-se inviável, devido à queda no número de currículos recuperados, que caiu de mais de 80% para menos de 10% após o descarte de um único termo. O descarte da palavras-chave, implica em uma uma perda do poder de recuperação do VSM, logo implica também em perda do poder de representação. Isso pode ser um problema

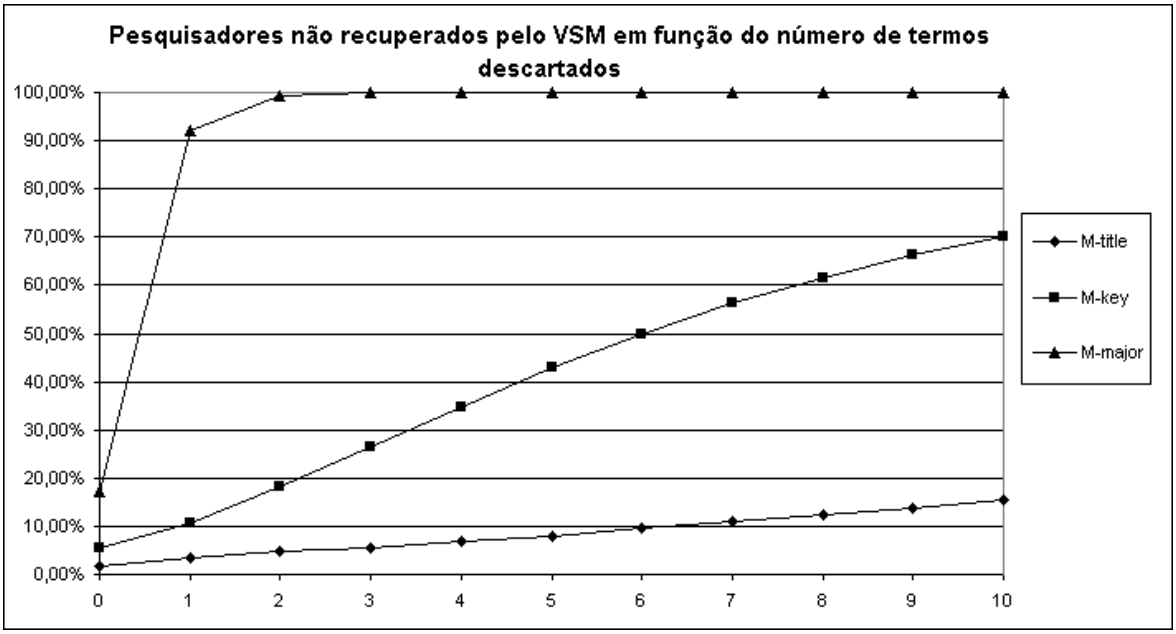

Figura 6.1: Impacto do descarte de termos na recuperação de currículos

para identificação de publicações inovadoras, que introduzam termos novos nos currículos, pois a essas características não seriam captadas de imediado pelo VSM, mas somente após o número de referências aos termos significativos ultrapassar o limite de corte usado para descarte de termos de baixa frequência.

A figura 6.1 apresenta a comparação do percentual do número de pesquisadores não considerados pelos modelos VSM, construídos com bases reduzidas através do aumento da frequência para descarte de termos. O descarte de termos no modelo M-major (última formação do pesquisador) mostrou-se inviável, devido ao crescimento vertiginoso do número de currículos não recuperados que passou de 17,24% para 92,22% após descarte de termos com frequência unitária. Note que 17,24% dos currículos na base de pesquisadores não apresentam título ou especialidade da última formação (frequência nula). Uma possível explicação para esse fato pode ser a ausência da informação do título e especialidade associados à formação de pós-doutorado. Nos estudos seguintes não foi considerado descarte de nenhum termo extraído da titulação do pesquisador para a construção do modelo M-major. A determinação da frequência máxima de termos para descarte com os modelos M-key

(palavras-chave da produção do pesquisador) e M-title (termos extraídos da produção do pesquisador e da sua especialidade) foi baseada em um estudo de clusterização dos pesquisadores com base em similaridade de perfil da produção, medida com esses modelos.

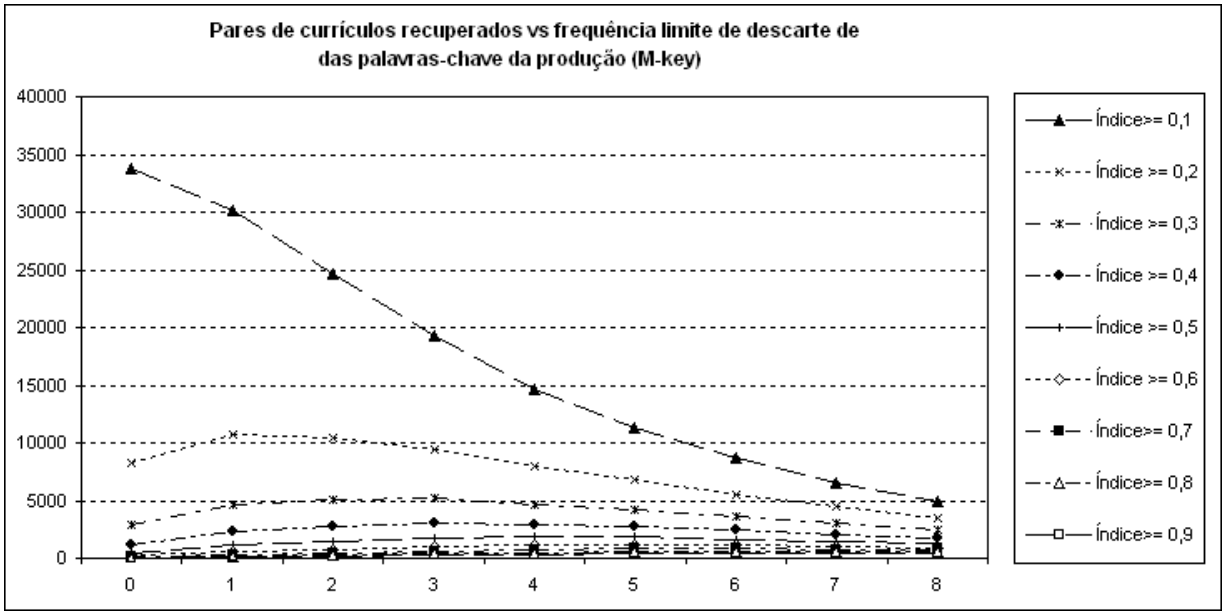

Figura 6.2: Pares de pesquisadores recuperados vs frequência de descarte (M-key)

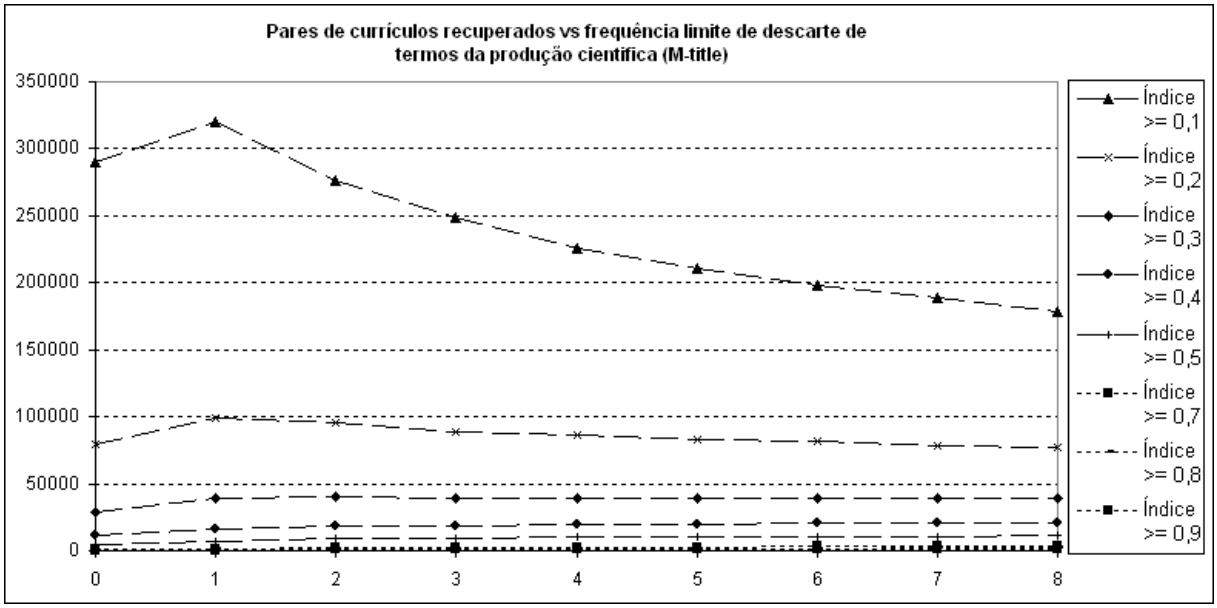

Figura 6.3: Pares de pesquisadores recuperados vs frequência de descarte (M-title)

As figuras 6.2 (M-key) e 6.3 (M-title) apresentam a evolução dos aglomerados de pesquisadores com produção similar com o aumento da frequência máxima para descarte de termos. Deseja-se aglomerados mais homogêneos, com alto índice de similaridade, pois admite-se que o *ad-hoc* terá melhores condições para julgar uma proposta se for ativo, mensurado pela produção recente, nos temas, subárea ou especialidade do proponente da proposta. Para similaridades muito baixas entre os pesquisadores (maior ou igual a 0,1), há uma acentuada redução no número de duplas de pesquisadores, recuperadas via o modelo M-key, com o descarte de termos, variando de 33.710 (sem descarte) a 4.857 (descarte de termos de frequência até 8). O número de 33.710 implica elevada dimensionalidade das bases do modelo VSM e afeta diretamente o cálculo do cosseno entre os vetores que representam dois pesquisadores, nessa base. Esse cosseno é utilizado para compor os elementos das matrizes de similaridade a serem criadas. Além desse fato, o comportamento do aglomerado para esse índice de similaridade (>= 0,1) foi considerado muito atípico em relação aos comportamentos das curvas associadas aos demais índices de similaridades. A curva associada ao índice de similaridade maior ou igual a 0,2 foi considerada mais representativa e escolhida para análise da frequência máxima de corte unitária para os modelos M-key e M-title (figuras 6.2 e 6.3).

Em resumo, os experimentos realizados sugerem que o descarte de termos pode ser aplicado a termos com frequência um ou, no máximo, dois para os modelos VSM construídos com as palavras-chave e com termos da produção científica. Nenhum descarte de termos de baixa frequência pode ser utilizado no modelo VSM da última formação do pesquisador.

Uma outra hipótese estuda foi o uso de vocabulário estruturado para reduzir dimensão do espaço vetorial dos modelos VSM. Face à indisponibilidade de tais vocabulários para as diversas áreas do conhecimento, os estudos paramétricos realizados focaram apenas o uso do DeCS – Descritores em Ciências da Saúde da BIREME – Centro Latino-Americano e do Caribe de Informação em Ciências da

Saúde, ex-Biblioteca Regional de Medicina, para análise de propostas de projetos da área de Ciências da Saúde. Os estudos experimentais indicaram que o uso desse vocabulário estruturado implicou em apenas 5% na redução da dimensionalidade da base VSM para essa área de conhecimento, considerado insuficiente, face ao aumento do tempo de processamento de busca de termos equivalentes de mais alta ordem (sinônimos).

6.3 Análise da Performance da Abordagem Proposta

Os resultados obtidos aplicando a metodologia proposta são comparados, quantitativamente, com as recomendações de consultores aceitas pelo CNPq para as propostas submetidas e avaliadas pelo CNPq. A hipótese subliminar nessa avaliação é que a abordagem adotada pelo sistema atual é adequada. Os índices de performance esperados para a abordagem proposta tendem a ser piores pois podem ser, no máximo, iguais aos obtidos com o sistema atual ou com a indicação direta de consultor *ad-hoc* feita pela equipe técnica do CNPq. Para avaliar a hipótese subliminar de adequabilidade da abordagem atual foi realizado um estudo comparativo qualitativo dos índices de similaridades entre os perfis dos currículos dos *ad-hoc* que emitiram pareceres e os perfis dos projetos por eles analisados.

O sistema atual e a abordagem proposta utilizam o valor 10 para o parâmetro número máximo de sugestões de consultor por proposta. Como a equipe técnica do CNPq indica, em geral, dois *ad-hoc* por proposta, índices tendem a ser limitados a 20%. Para permitir uma análise de sensibilidade dos índices ao parâmetro citado, foram plotados gráficos nos quais os índices de performance são calculados considerando a lista de *ad-hoc* recomendados, variando de um até dez ad-hoc.

As figuras 6.4, 6.5 e 6.6 apresentam uma comparação quantitativa entre a abordagem proposta e a baseada no sistema atual, com o uso dos índices de desempen-

hos clássicos (Rijsbergen, 1979) para sistemas de recomendação, adaptados para o domínio em questão:

$$recall = \frac{CRI}{CI}$$

$$precision = \frac{CRI}{CR}$$

$$F\text{-}measure = \frac{2 \cdot precision \cdot recall}{(precision + recall)}$$

onde:

- CRI é o número de consultores recomendados (sistema atual ou abordagem proposta) indicados pelo CNPq,
- CR é o número de consultores recomendados (sistema atual ou abordagem proposta),
- CI é o número de consultores indicados pelo CNPq (a partir de recomendações ou diretamente pela equipe técnica). A indicação do consultor pelo CNPq foi considerada como medida de relevância.

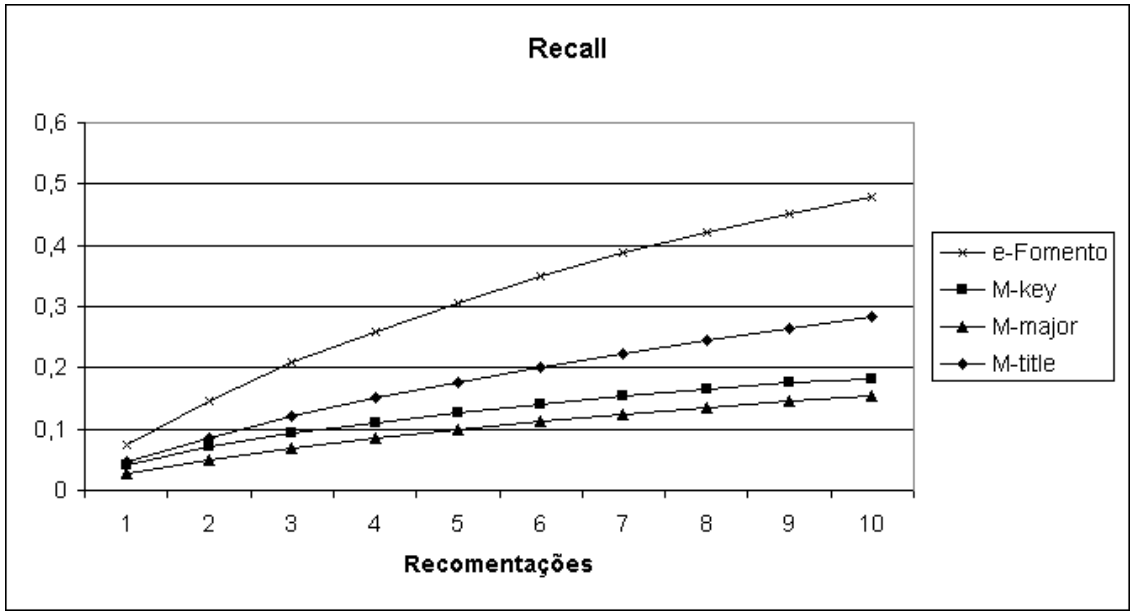

Figura 6.4: *Recall* para as abordagens atual e proposta

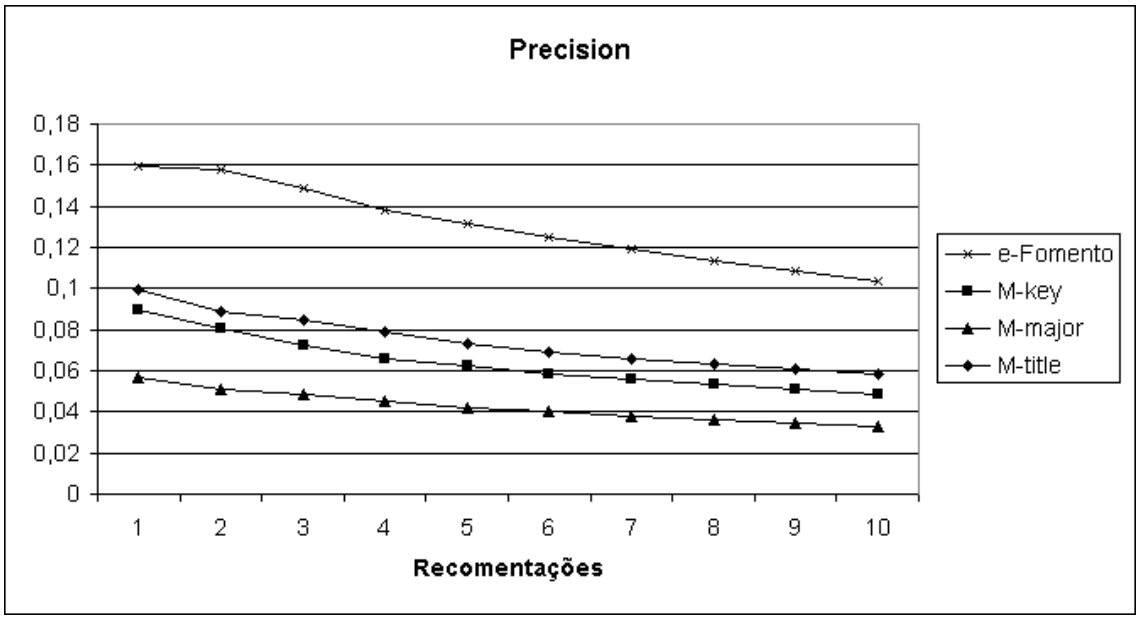

Figura 6.5: *Precision* para as abordagens atual e proposta

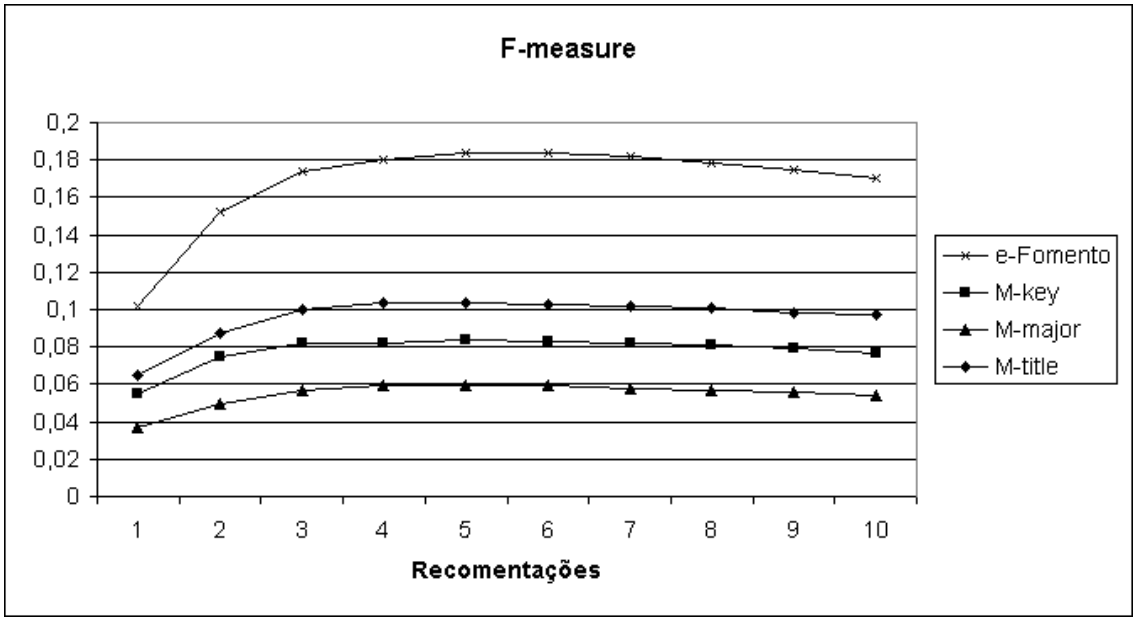

Figura 6.6: *F-Measure* para as abordagens atual e proposta

Os valores para esses índices para a abordagem atual (sistema atual e indicação direta) são superiores aos obtidos para os três métodos da abordagem proposta: M-key (perfis compostos com palavras-chave da produção nos últimos 5 anos), M-title (perfis compostos com termos extraídos das palavras-chaves, especialidade da área e títulos da produção nos últimos 5 anos), e M-major (perfis compostos com termos

retirados do título da última formação e da especialização do pesquisador).

Nos cálculos dos índices foram utilizados quantitativos baseados no número de consultores recomendados e no número de consultores recomendados ou não, mas que foram indicados pela equipe técnica do CNPq e que emitiram pareceres, pois, na abordagem atual, não existem dados disponíveis que permitam concluir que consultores são mais adequados para avaliar cada uma das propostas.

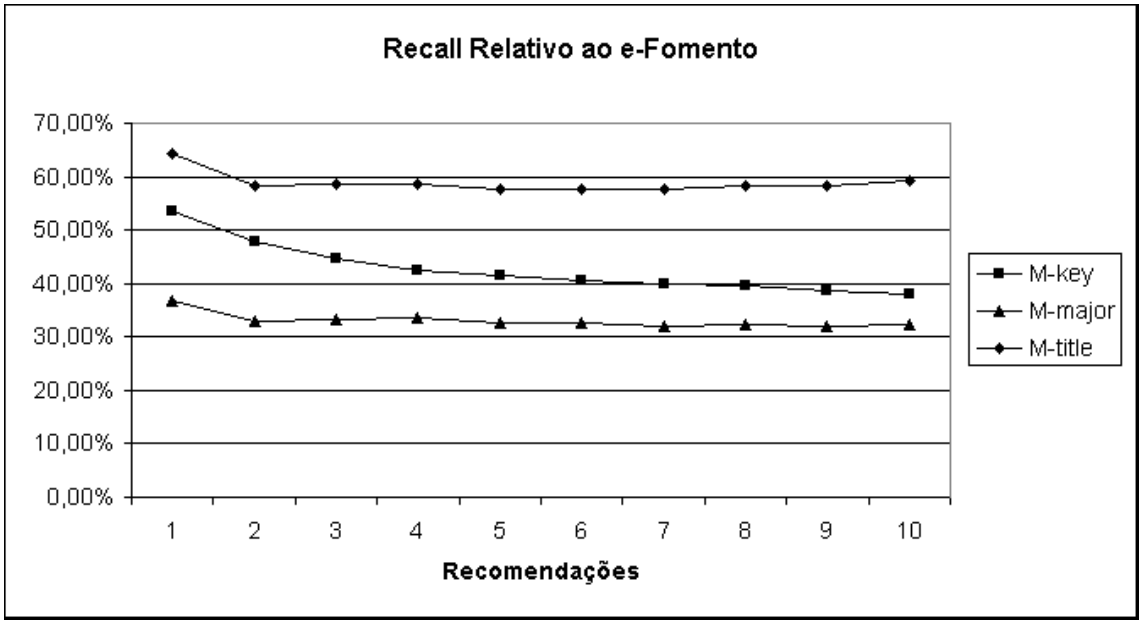

Figura 6.7: *Recall* da abordagem proposta em relação ao sistema atual

Os dados nas figuras 6.4, 6.5 e 6.6 apresentam a evolução dos índices de desempenho do sistema atual e dos três modelos da abordagem proposta. Foram calculados os índices de desempenho para conjuntos de recomendações variando de um a dez consultores recomendados por proposta. Esses dados sugerem que a abordagem atual é melhor do que a abordagem proposta.

As figuras 6.7, 6.8 e 6.9 apresentam os índices de performance relativos aos índices de performance obtidos com a abordagem atual (sistema atual, mais indicações realizadas pela equipe técnica) para conjuntos de recomendações variando de um a dez consultores recomendados por propostas. A análise desses quadros evidencia que a construção de índices de similaridades entre os perfis dos consultores e

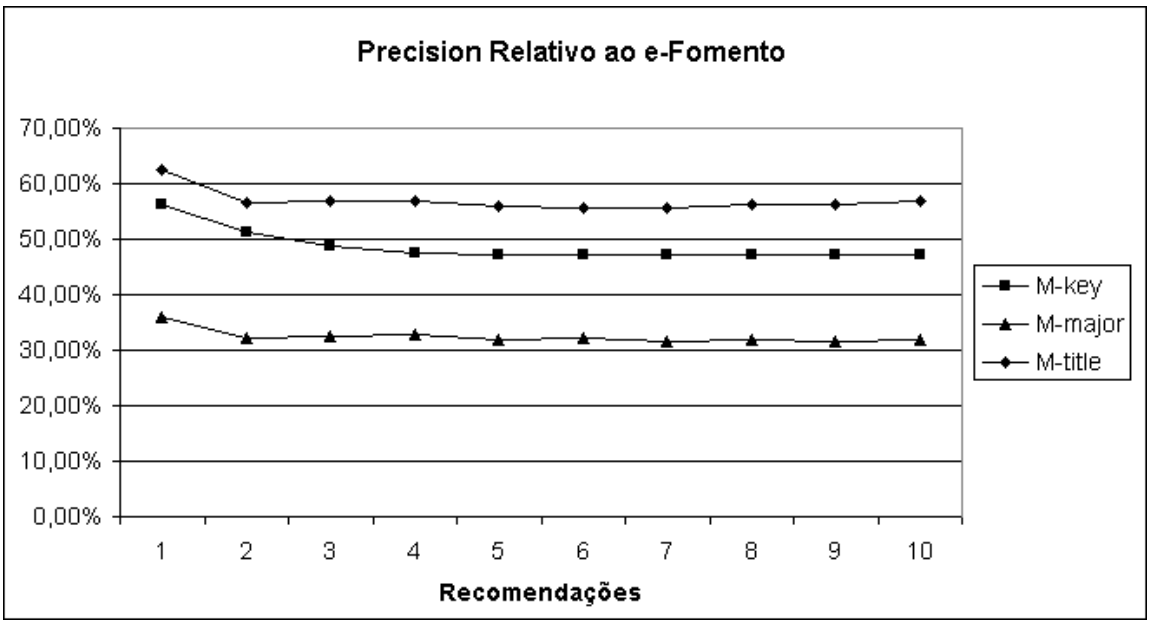

Figura 6.8: *Precision* da abordagem proposta em relação ao sistema atual

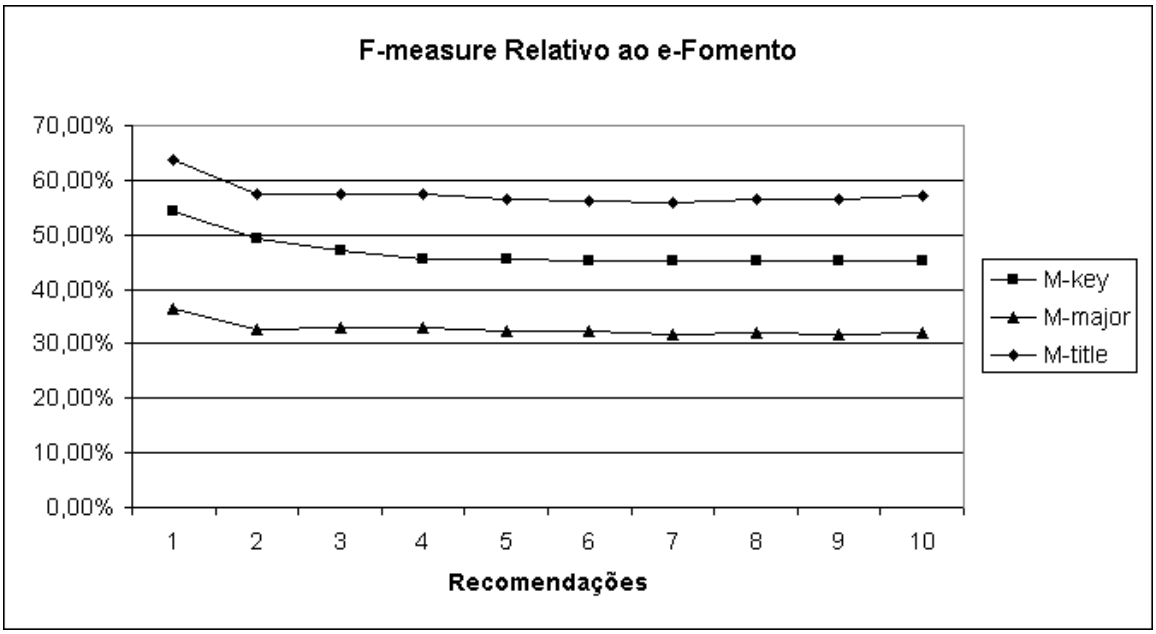

Figura 6.9: *F-Measure* da abordagem proposta em relação ao sistema atual

propostas baseados no modelo M-title (considerando palavras-chave, títulos da produção nos últimos 5 anos, e especialidades da subáreas em que se enquadraram) é a que melhor reproduz os índices obtidos com a abordagem atual do CNPq. Isso contraria a ideia de que as palavras-chave comporiam um indexador melhor, uma vez que as mesmas são escolhidas pelos próprios autores como descritores de suas

produções científicas e tecnológicas.

Para estudar a hipótese subliminar de que a abordagem atual é correta, foram realizados estudos da similaridade entre os perfis dos consultores *ad-hoc* pareceristas e os perfis das propostas de projetos que analisaram. A tabela 6.4 apresenta uma comparação dos valores médios dos coeficientes de similaridade para os consultores recomendados indicados que emitiram pareceres na abordagem atual e na abordagem proposta. Esses valores foram calculados para cada um dos três modelos apresentados. Os coeficientes de similaridade entre os perfis dos consultores e das propostas de projetos por eles avaliadas, em todos os modelos da abordagem proposta (M-key, M-title e M-major), com ou sem descarte de termos de frequência unitária nos currículos, são superiores aos coeficientes de similaridade obtidos por meio da abordagem atual. Na abordagem M-major não houve descarte de termos pois se considerou apenas a última formação do pesquisador, como já mencionado na descrição do estudo de casos realizado.

	Abordagem atual sem descarte de termos	**Abordagem proposta sem descarte de termos**	**Abordagem atual com descarte de termos**	**Abordagem proposta com descarte de termos**
Última formação	0,471	0,540		
Palavras-chave	0,473	0,538	0,473	0,557
Produção científica	0,482	0,554	0,482	0,554

Tabela 6.4: Comparação dos *scores* da abordagem atual X abordagem proposta

A tabela 6.5 apresenta a diferença percentual entre os índices de similaridade médios, relativas aos índices obtidos com a abordagem atual, sugerindo que as recomendações da abordagem proposta são qualitativamente superiores às da abordagem atual em qualquer dos três modelos testados.

	Sem descarte de termos	**Com descarte de termos**
Última formação	14,76%	
Palavras-chave	13,81%	17,82%
Produção científica	14,74%	14,98%
Média	**14,43%**	**16,40%**

Tabela 6.5: Comparação % dos *scores* da abordagem atual X abordagem proposta

6.4 Dificuldades encontrados

As principais dificuldades à realização deste trabalho:

- volume de dados elevado;
- tempo de processamento excessivamente longo;
- limitações ao uso de tempo de processamento e de uso de espaço nos servidores de banco de dados do CNPq;
- diversidade línguas presentes nos termos (palavras-chave, títulos, resumos, etc.) da produção científica dos consultores;
- ausência de um dicionário de termos, ou tesauro, para padronização das palavras-chave no cadastro de currículos e de propostas, resultando em dispersão de termos por problemas de grafia, abreviação e sinonímia;
- as citações bibliográficas registradas de formas variadas, com ocorrência de cadastros sem integridade referencial – o autor pode usar mais de um nome em suas próprias publicações, e pode ser citado utilizando outras variações diferentes daquela pretendida pelo autor;
- o uso de mecanismos de busca por aproximação nas citações bibliográficas e registros de orientação de alunos tornou o processamento desses dados excessivamente lento; e
- não foi possível usar o parâmetro dependente da carga de trabalho atribuída aos consultores.

Capítulo 7

Conclusão e desenvolvimentos futuros

Dos consultores indicados pelos técnicos do CNPq, 67,47% foram recomendados pelo sistema atual de recomendação de consultores *ad-ho* mas 9,82% (28,53% deles) não emitem o parecer. Portanto o desempenho real médio do sistema atual é de apenas 54,65% (tabela 4.1). Para os demais 32,53% do total de consultores indicados diretamente pelos técnicos, 6,59% (20,26% deles) não emitem o parecer, o que corresponde a um desempenho médio real de 25,94%. As razões mais frequentes para a área técnica do CNPq rejeitar uma recomendação do sistema atual são: a) o consultor recomendado já pode ter sido indicado para o número máximo de propostas por consultor (tabela 6.2, valor em uso no sistema atual é 4), b) o sistema pode não ter recomendado nenhum *ad-hoc* por não ter encontrado nenhum consultor que atue na área do conhecimento da proposta e que não tenha restrição para ser recomendado; e c) o técnico pode não ter concordado com as recomendações do sistema. Em geral, há uma tendência de que a área técnica acate as recomendações do sistema seguindo a ordem em que são apresentadas. Portanto, esses consultores tendem a receber o máximo de propostas para análise permitida pelo sistema. Como o sistema atual analisa perfis pela área de conhecimento da

proposta pode ocorrer casos em que consultores que atuem em mais de uma área não sejam localizados. O desempenho final, avaliado em função do número de pareceres emitidos, das indicações realizadas com base nas recomendações (81%) e das indicações realizadas sem recomendação automática (79,75%) são equivalentes.

Na metodologia de avaliação se adotou a hipótese de que consultores relevantes são os indicados pela área técnica do CNPq, quer com base no sistema atual ou não. Do ponto de vista quantitativo – mensurado com os índices de performance – a abordagem proposta apresentou desempenho inferior ao desempenho do sistema atual (figuras 6.4 a 6.9), sendo o modelo M-title o que apresentou desempenho mais alto em relação aos modelos M-key e M-major. Essa análise quantitativa assume a hipótese subliminar de que a abordagem atual do CNPq está correta, pois considera a indicação de consultor pelo CNPq como medida de relevância no cálculo dos índices de performance *precison*, *recall* e *F-measure*. Como uma forma de avaliar a veracidade dessa hipótese, foi realizado um estudo qualitativo dos índices de similaridade entre consultor e proposta a ser avaliada por ele. Do ponto de vista dessa análise qualitativa, a abordagem proposta recomendou consultores com perfis mais similares aos das propostas que irão analisar, portanto apresentando desempenho qualitativo superior ao obtido por meio do sistema tual, para qualquer um dos modelos M-key, M-title ou M-major, independente de ter havido ou não descarte de termos de baixa frequência (tabelas 6.4 e 6.5). O descarte de termos de baixa frequência mostrou-se eficaz na redução dimensional da base de vetores VSM, sem degradar o modelo proposto para o cálculo da similaridade entre os perfis envolvidos.

Dentre os três modelos estudos, o M-key com descarte de termos – construção de índices de similaridades baseados em palavras-chave extraídas da produção dos últimos 5 anos com frequência superior a um e todas as palavras-chave contidas na proposta do projeto e currículo do consultor – apresentou melhor desempenho qualitativo (maior similaridade entre perfil do consultor e da proposta) e computacional

espaço com dimensões reduzidas (tabela 6.3).

As metodologias proposta e a atual selecionam consultores que atuam na mesma área de conhecimento da proposta de projeto e atribuem peso diferente de zero (figura 6.2) para subárea ou especialidade apenas para aqueles que são da mesma subárea ou especialidade da proposta. Como trabalho futuro, a recomendação de consultores de áreas distintas da área de conhecimento da proposta do projeto facilitará a análise de propostas com temas multidisciplinares. Para tanto seria adequada a construção de uma ontologia de conceitos baseados nos termos conforme ocorram nos diversos níveis das áreas do conhecimento.

O uso de vocabulário estruturado por área do conhecimento pode ser de grande utilidade para redução da cardinalidade da base do VSM, mantendo o poder de expressão do modelo. Essa alternativa foi avaliada – na área de Ciências da Saúde com a utilização do DeCS/BIREME – mas não foi usada nesse projeto, pois o cadastramento dos currículos e das propostas foram realizados sem a aplicação de tais recursos e seu uso no estudo de caso indicou um ganho de cerca de 5% na dimensão das bases, mas com aumento considerável no esforço computacional. Esse mecanismo será mais útil se for usado desde a entrada de dados do Currículo Lattes e do formulário de proposta, pois reduz os erros de digitação e o uso de sinônimos.

7.1 Estudos e desenvolvimento futuro

Este trabalho aborda apenas uma pequena fração das necessidades envolvidas na recomendação de consultores *ad-hoc*, mesmo que essa necessidade ficasse restrita ao escopo do CNPq. Dada a importância do tema e a possibilidade da exploração da abordagem proposta em outros contextos, algumas alternativas e de estudos e desenvolvimentos futuros devem ser considerados:

- mecanismos para gerar recomendações de consultores *ad-hoc* fora da área de conhecimento da proposta de projeto;

- ontologia de áreas de conhecimento combinando a tabela de áreas do conhecimento com as áreas cadastradas nos currículos dos pesquisadores;
- suporte para os diversos idiomas presentes nos currículos dos pesquisadores;
- uso de um dicionário de termos, ou um tesauro, no currículo e no formulário eletrônico de propostas para melhorar a representação pelas palavras-chave;
- avaliação das recomendações automáticas pelos técnicos que usam o sistema, com retroalimentação para novas recomendações; e
- uso da justificativa de solicitação de dispensa de emissão de parecer fornecida pelo consultor *ad-hoc* indicado como retroalimentação do sistema de recomendação automática.

Referências Bibliográficas

[Birukov et al., 2005] Birukov, A., Blanzieri, E., and Giorgini, P. (2005). Implicit: A recommender system that uses implicit knowledge to produce suggestions. In *Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05)*, pages 418–624, Edinburgh, Scotland. University of Trento. 9, 41, 42

[Borko and Bernick, 1963] Borko, H. and Bernick, M. (1963). Automatic document classification. *Journal of the ACM*, 10(2):151–162. 27

[Caid and Carleto, 2003] Caid, W. R. and Carleto, J. L. (2003). Context vector-based text retrieval. Site acessado em 21/11/2007. 21

[Cazella and Alvares, 2005] Cazella, S. C. and Alvares, L. O. C. (2005). Combining data mining technique and users' relevance opinion to build an efficient recommender system. *Revista Tecnologia da Informação*, 5(1):9–20. 42, 44

[CNPq, 2007] CNPq (2007). Site oficial do conselho nacional de desenvolvimento científico e tecnológico - cnpq. Site acessado em 01/11/2007. 4

[Florid, 2003] Florid, L. (2003). *The Blackwell Guide to the Philosophy of Computing and Information*. Oxford University Press, New York, USA. 31

[Gonçalves and Souza, 1977] Gonçalves, A. and Souza, R. (1977). *Introdução à Àlgebra Linear*. Editora Blücher Ltda, São Paulo, SP. 16

[Han and Karypis, 2005] Han, E.-H. and Karypis, G. (2005). Feature-based recommendation system. pages 446–452, Bremen, Germany. 1, 9

[Ikehara et al., 2001] Ikehara, S., Murakami, J., Kimoto, Y., and Araki, T. (2001). Vector space model based on semantic attributes of words. 19, 24, 27

[Kuropka, 2003] Kuropka, D. (2003). *Modelle zur Repräsentation natürlichsprachlicher Dokumente*. Logos Verlag, Berlin, Germany. 31

[Lopes et al., 2006] Lopes, G. R., Souto, M. A. M., and de Oliveira, J. P. M. (2006). Sistema de recomendação para bibliotecas digitais sob a perspectiva da web semântica. *II Workshop de Bibliotecas Digitais, WDL; SBBD/SBES*, pages 21–30. 45

[Manning et al., 2008] Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, USA. 16

[Mendes et al., 2002] Mendes, C. A., de Moura, E. S., and Ziviani, N. (2002). Expansão de consultas utilizando indexação semântica latente. pages 166–180. UFRGS. 16

[Monteiro, 1974] Monteiro, J. L. H. (1974). *Elementos de Álgebra*. Livros Técnicos e Científicos Editora SA, Rio de Janeiro, RJ. 16

[Oliveira et al., 2007] Oliveira, E., Ciarelli, P. M., Santos, M. H., and da Costa, B. O. (2007). An adaptive recommendation system without explicit acquisition of user relevance feedback. *Revista Brasileira de Biblioteconomia e Documentação*, 3(1):73–98. 25

[Polyvyanyy and Kuropka, 2007] Polyvyanyy, C. and Kuropka, D. (2007). A quantitative evaluation of the enhanced topic-based vector space model. Technical Report 19, Hasso Plattner Insitute, Berlin, Germany. 20, 22, 26, 27, 30, 31, 32, 33, 35

[Poncelet et al., 2008] Poncelet, P., Teisseire, M., and Masseglia, F. (2008). *Data Mining Patterns: New Methods and Applications*. Information science reference, Hershey, New York. 85

[Porter, 2006] Porter, J. (2006). Watch and learn: How recommendation systems are redefining the web. Sítio da Internet acessado em 05/12/2007. 11

[Recio-García et al., 2008] Recio-García, J. A., Díaz-Agudo, B., and González-Calero, P. (2008). jcolibri 2 tutorial – case-base reasoning framework. 29

[Salton et al., 1975] Salton, G. M., Wong, A., and Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620. 26

[Shahabi and Chen, 2003] Shahabi, C. and Chen, Y.-S. (2003). An adaptive recommendation system without explicit acquisition of user relevance feedback. *Distributed and Parallel Databases*, 14(2):173–192. 3, 14, 39, 40, 41

[van Rijsbergen B, 1979] van Rijsbergen B, C. J. (1979). Information retrieval. Site acessado em 06/08/2007. 36