UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

GISELI RABELLO LOPES

Sistema de Recomendação para Bibliotecas Digitais sob a Perspectiva da Web Semântica

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Prof. Dr. José Palazzo Moreira de Oliveira
Orientador

Profa. Dra. Maria Aparecida Martins Souto
Co-orientadora

Porto Alegre, fevereiro de 2007.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Lopes, Giseli Rabello

Sistema de Recomendação para Bibliotecas Digitais sob a Perspectiva da Web Semântica / Giseli Rabello Lopes – Porto Alegre: Programa de Pós-Graduação em Computação, 2007.

69 f.:il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2007. Orientador: José Palazzo Moreira de Oliveira; Co-orientadora: Maria Aparecida Martins Souto.

1.Sistemas de Recomendação. 2.Bibliotecas Digitais. 3.OAI. 4.Personalização da Informação. 5.Modelo Vetorial. 6.Provedor de Serviços. 7.Web Semântica. I. Oliveira, José Palazzo Moreira de. II. Souto, Maria Aparecida Martins Souto. III. Título.

“Descobri como é bom chegar quando se tem paciência. E para se chegar, onde quer que seja, aprendi que não é preciso dominar a força, mas a razão. É preciso, antes de mais nada, querer.”

— AMYR KLINK

AGRADECIMENTOS

Meus sinceros agradecimentos a todos que contribuíram para o desenvolvimento deste trabalho. Em especial, gostaria de agradecer:

Ao meu orientador, Prof. Dr. José Palazzo Moreira de Oliveira, por todos os valorosos conhecimentos e experiências transmitidos. Agradeço muito pela orientação, pelas oportunidades e pela atenção que me foram dispensadas ao longo de todo o curso de mestrado, as quais foram de fundamental importância para a realização deste trabalho.

À Profa. Dra. Maria Aparecida Martins Souto, que aceitou ser minha co-orientadora e sempre esteve disposta a ajudar, pela atenção dispensada, por seus comentários e sugestões sempre úteis que contribuíram imensamente para o desenvolvimento deste trabalho.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo suporte financeiro, durante todo o mestrado, que permitiu a realização deste trabalho com dedicação exclusiva.

Ao Instituto de Informática da UFRGS por toda a infra-estrutura disponibilizada e a seus profissionais; aos professores do PPGC pelos ensinamentos transmitidos nas disciplinas por eles ministradas e aos funcionários sempre solícitos e prestativos.

Aos professores e alunos dos grupos de Modelagem Conceitual e Adaptabilidade e Banco de Dados, pelas valiosas sugestões durante as reuniões do grupo de pesquisa, pela troca de conhecimentos e experiências.

Aos professores e alunos do Instituto de Informática da UFRGS que participaram da avaliação experimental apresentada neste trabalho, por se disporem a colaborar com a pesquisa realizada, pela paciência em avaliar as recomendações recebidas e pelos comentários elaborados que foram de grande valia nesta etapa.

Ao Prof. Dr. Leandro Krug Wives pela valiosa colaboração em co-autoria de artigo e pelas importantes sugestões para a avaliação dos resultados obtidos por este trabalho; e, à Profa. Dra. Viviane Moreira Orengo, que também contribuiu significativamente com sugestões para a avaliação e análise dos resultados alcançados.

Aos membros da banca de defesa desta dissertação: Prof. Dr. Mario Lemes Proença Jr. (UEL), Prof. Dr. Carlos Alberto Heuser (UFRGS) e Profa. Dra. Viviane Moreira Orengo (UFRGS); por terem aceitado o convite, pelas importantes sugestões e correções que contribuíram no aprimoramento do texto final deste trabalho e pelo incentivo para o seguimento da pesquisa realizada.

A todos os meus amigos que acompanharam, de perto ou de longe, a realização deste trabalho e que torcem para que o mesmo seja concluído com êxito. Aos amigos e

colegas de grupo de pesquisa na UFRGS: Alexander Vinson, Eduardo Borges, Gabriel Simões, Marcos Nunes, Mariusa Warpechowski e Sérgio Mergen, que propiciaram um ambiente de integração e troca de experiências, que partilharam momentos de seriedade e descontração durante esses dois anos. Ao Alexander, que também foi companheiro de trabalhos e disciplinas durante o mestrado, agradeço pela amizade e por estar sempre disposto a colaborar tanto para discutir idéias, ler rascunhos ou trocar experiências.

Aos meus pais por todo amor, companheirismo e confiança que sempre me dispensaram, agradeço as importantes lições de vida que me foram transmitidas por eles ao longo da minha existência e o apoio incondicional que sempre me oferecem. A eles, que tantas vezes de longe estiveram perto e me ajudaram muito mais do que imaginam. A distância que nos separou só me fez ver ainda mais que posso contar sempre com eles. Muito obrigada, se pude chegar até aqui, foi graças ao incentivo de vocês.

À minha irmã pelo carinho e torcida mesmo à distância.

Ao meu namorado, Daniel da Costa Mendes, por estar sempre me apoiando e incentivando, por ser conforto também nos momentos de adversidade. Agradeço pela atenção, confiança, amor e carinho dispensados, e por entender minhas ausências.

Por fim, agradeço imensamente a Deus por ter me dado forças em mais esta etapa de minha vida e por ter me propiciado mais esta oportunidade de crescimento pessoal e profissional.

> “Deus é o nosso refúgio e fortaleza, socorro bem presente nas tribulações. Portanto, não temeremos ainda que a terra se transtorne e os montes se abalem nas profundezas dos mares; ainda que as águas rujam e espumem e na sua fúria os montes se estremeçam.”
>
> (Salmos 46,1-3)

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

ACM	Association for Computing Machinery
API	Application Programming Interface
ARIADNE	Annotatable Retrieval of Information And Database Navigation Environment
ARP	Adaptative Recommendation Project
BDBComp	Biblioteca Digital Brasileira de Computação
CITIDEL	Computing and Information Technology Interactive Digital Educational Library
CLEF	Cross Language Evaluation Forum
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CONSCIENTIAS	Comunidade para Ontologias em Ciência, Tecnologia e Informações de Aperfeiçoamento de Nível Superior
CTInfo	Comitê da Área de Tecnologia da Informação
CV	Curriculum Vitae
DB	DataBase
DBLP	Digital Bibliography & Library Project
DC	Dublin Core
DCMES	Dublin Core Metadata Element Set
DCMI	Dublin Core Metadata Initiative
EAD	Educação a Distância
ER	Entidade-Relacionamento
FAPERGS	Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul
HTTP	Hypertext Transfer Protocol
IDF	Inverse Document Frequency
IR	Information Retrieval
ISBN	International Standard Book Number
JEMS	Journal and Event Management System
LPML	Linguagem de Marcação da Plataforma Lattes

NCSA	National Center for Supercomputing Applications
OAI	Open Archives Initiative
OAI-PMH	Open Archives Initiative - Protocol for Metadata Harvesting
OCLC	Online Computer Library Center
PhD	Doctor of Philosophy
PHP	Hypertext Preprocessor
PPGC	Programa de Pós-Graduação em Computação
PRONEX	Programa de Apoio a Núcleos de Excelência
SAX	Simple API for XML
SBBD	Simpósio Brasileiro de Banco de Dados
SBC	Sociedade Brasileira de Computação
SQL	Structured Query Language
TF	Term Frequency
TREC	Text REtrieval Conference
UEL	Universidade Estadual de Londrina
UFMG	Universidade Federal de Minas Gerais
UFRGS	Universidade Federal do Rio Grande do Sul
URL	Uniform Resource Locator
VSM	Vector Space Model
W3C	World Wide Web Consortium
WWW	World Wide Web
XML	eXtensible Markup Language

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Atualmente, pesquisadores e acadêmicos têm beneficiado-se muito com o crescimento acelerado das tecnologias Web, pois os resultados de pesquisa podem ser publicados e acessados eletronicamente tão logo a mesma tenha sido realizada. Esta possibilidade é vantajosa na medida em que minimiza as barreiras de tempo e espaço associadas à publicação tradicional. Neste contexto, surgem as Bibliotecas Digitais como repositórios de dados que, além dos documentos digitais propriamente ditos, ou de apontadores para estes documentos, armazenam os metadados associados. Para permitir que diferentes Bibliotecas Digitais possam interoperar surgiu a *Open Archives Initiative* (OAI) e, para resolver a questão da padronização dos metadados utilizados pelos repositórios, foi criado o formato Dublin Core (DC).

Por outro lado, a enorme quantidade de documentos digitais disponíveis na Web tem causado o fenômeno conhecido como "sobrecarga de informação". Com o objetivo de suprir esta dificuldade, Sistemas de Recomendação têm sido propostos e desenvolvidos. Estes sistemas visam prover uma interface alternativa para tecnologias de filtragem e recuperação de informações, tendo como foco a predição daqueles itens ou partes da informação que o usuário acharia interessante e útil. Portanto, os Sistemas de Recomendação atuam baseados em personalização da informação sendo que as predições geralmente são realizadas utilizando-se um perfil de cada usuário. A personalização está relacionada com o modo pelo qual a informação e serviços podem ser ajustados às necessidades específicas de um usuário ou comunidade.

Esta dissertação descreve um Sistema de Recomendação de artigos científicos, armazenados em bibliotecas digitais. Este sistema é dirigido à comunidade científica da área da Ciência da Computação. Tecnologicamente, o sistema proposto foi desenvolvido sob a perspectiva da Web Semântica, à medida que faz uso de suas tecnologias emergentes tais como: uso de metadados padrão para a descrição de documentos - Dublin Core, uso do padrão XML para a descrição do perfil do usuário - Currículo Lattes, e provedores de serviços e de dados (OAI) envolvidos no processo de geração das recomendações. Este trabalho ainda apresenta e discute alguns resultados de experimentos baseados em avaliações quantitativas e qualitativas de recomendações geradas pelo sistema.

Palavras-Chave: Sistemas de Recomendação, Bibliotecas Digitais, OAI, Personalização da Informação, Modelo Vetorial, Provedor de Serviços, Web Semântica.

A Recommender System to Digital Libraries under Semantic Web Perspective

ABSTRACT

Currently, researchers and academics have been benefited by the expressive growth of web technologies, due to the possibility of publishing and accessing research results as soon as they are achieved. This possibility is advantageous as it minimizes the time and space barriers that traditional publications present. In this context, Digital Libraries emerged as data repositories that, beyond digital documents or links to them, store associated metadata. To allow the interoperability among different Digital Libraries, the Open Archives Initiative (OAI) was defined and, to solve the problem of metadata standardization, the Dublin Core standard (DC) was created.

On the other hand, the great amount of available digital documents in the Web has caused the phenomenon known as "information overload". In order to avoid this difficulty, Recommender Systems have been proposed and developed. These systems intend to provide an alternative interface for information filtering and retrieval technologies, focusing on the prediction of items or information parts that are interesting and useful for the user. Therefore, Recommender Systems act based on information personalization, and the predictions are generally generated using each user's profile. The personalization is related to the way the information and the provided services can be adjusted to the specific necessities of a user or community.

This dissertation describes a Recommender System for scientific articles stored in digital libraries. This system is geared towards the Computer Science scientific community. Technologically, the proposed system was developed under the Semantic Web perspective, as it explores its emergent technologies such as: use of standard metadata for document description - Dublin Core, use of the XML standard for users' profile description - Lattes Curriculum Vitae, and services and data providers (OAI) involved on the recommendations generation process. In addition, this work presents and discusses some experimental results; the experiments are based on quantitative and qualitative evaluations of recommendations generated by the system.

Keywords: Recommender Systems, Digital Libraries, OAI, Information Personalization, Vector Space Model, Service Providers, Semantic Web.

1 INTRODUÇÃO

O expressivo crescimento das tecnologias Web tem beneficiado pesquisadores e acadêmicos. Nos dias atuais, as publicações de pesquisa podem ser acessadas eletronicamente tão logo elas tenham sido finalizadas e publicadas na Web. A principal vantagem da publicação aberta é a minimização das barreiras de tempo e espaço inerentes ao processo de publicação tradicional.

Neste contexto, surgem as Bibliotecas Digitais como repositórios de dados que, além dos documentos digitais propriamente ditos, ou de apontadores para estes documentos, armazenam os metadados associados. Muitos sistemas de Bibliotecas Digitais têm sido desenvolvidos, entre eles EPrints (GUTTERIDGE, 2002), DSpace (TANSLEY et al., 2003), Kepler (MALY et al., 2004) e CITIDEL (*Computing and Information Technology Interactive Digital Educational Library*) (CITIDEL, 2005). No Brasil deve ser citada a BDBComp (Biblioteca Digital Brasileira de Computação) (LAENDER; GONÇALVES; ROBERTO, 2004).

Por outro lado, a enorme quantidade de documentos digitais disponíveis na Web tem causado o fenômeno conhecido como "sobrecarga de informação" (*information overload*) que dificulta bastante os processos de busca *online* (HUANG et al., 2002) por parte dos usuários. Normalmente, usuários com diferentes níveis de conhecimento, experiência e interesse são igualmente providos com a mesma informação, em resposta a uma mesma consulta. Com o objetivo de suprir estas dificuldades, Sistemas de Recomendação para Bibliotecas Digitais têm sido propostos e desenvolvidos (HUANG et al., 2002; HWANG; HSIUNG; YANG, 2003; CALLAN et al., 2003). Além desses, citamos os projetos ARIADNE (ARIADNE, 2006), ResearchIndex (COSLEY; LAWRENCE; PENNOCK, 2002), CyberStacks (CYBERSTACKS, 2006) e ARP (ARP, 2006).

Os Sistemas de Recomendação atuam baseados em personalização da informação. A personalização está relacionada com o modo pelo qual a informação e serviços podem ser ajustados às necessidades específicas de um usuário ou comunidade (CALLAN et al., 2003; DOLOG; NEJDL, 2003). Esta funcionalidade pode ser obtida através da adaptação da apresentação, conteúdo e/ou serviços baseada na atividade da pessoa, bagagem cognitiva, histórico, necessidades de informação, localidade, etc.

O presente trabalho insere-se no contexto acima exposto. Especificamente, este trabalho apresenta um Sistema de Recomendação de artigos científicos, na área da Ciência da Computação, que estejam de acordo com os interesses do usuário identificados a partir de informações presentes em seu currículo Lattes. Sob o ponto de vista tecnológico, o sistema proposto foi desenvolvido sob a perspectiva da Web Semântica, à medida que faz uso de suas tecnologias emergentes tais como: uso de metadados padrão para a descrição de documentos - Dublin Core (DUBLIN, 2005), uso

de padrão XML para a descrição do perfil do usuário - currículo Lattes (Lattes-CNPq, 2005), e utilização de provedor de serviços e dados para gerar a recomendação.

Este trabalho está organizado da seguinte maneira:

O capítulo 2 apresenta o contexto tecnológico no qual o sistema de recomendação foi desenvolvido, discutindo assuntos como Web Semântica, Bibliotecas Digitais, padrão OAI e o formato Dublin Core.

O capítulo 3 apresenta as abordagens existentes em sistemas de recomendação, comentando brevemente alguns exemplos de sistemas existentes, bem como discute as decisões adotadas neste trabalho.

O capítulo 4 detalha o sistema de recomendação desenvolvido, bem como aspectos de sua implementação, incluindo: a descrição do perfil do usuário (currículo Lattes), a apresentação da arquitetura do sistema e do modelo de recomendação adotado.

O capítulo 5 descreve os experimentos de avaliação do sistema e apresenta a análise de alguns resultados importantes obtidos.

Por fim, o capítulo 6 apresenta algumas considerações finais, bem como apresenta os trabalhos futuros.

2 BIBLIOTECAS DIGITAIS E A WEB SEMÂNTICA

A Web Semântica (*Semantic Web*), idealizada por Berners-Lee (1999), é uma extensão da Web atual, na qual a informação é gerada, não somente para leitores humanos, mas também para processamento por máquinas, possibilitando serviços de informação inteligentes, *Web-sites* personalizados e máquinas de busca semanticamente enriquecidas. Para atingir esta meta, um dos importantes requisitos consiste em disponibilizar metadados de descrição dos recursos Web, tanto sob o ponto de vista dos conteúdos destes recursos, quanto sob o ponto de vista de suas funcionalidades.

No contexto das Bibliotecas Digitais, as tecnologias da Web Semântica têm um papel importante à medida que possibilitam acesso eficiente e inteligente aos documentos digitais na Web. O uso de padrões para a descrição dos objetos de informação baseados em metadados (meta-informações associadas) apresenta duas grandes vantagens: obtenção de maior eficiência computacional durante a colheita de informações; e possibilidade de se obter interoperabilidade entre as Bibliotecas Digitais. Assim, para permitir que diferentes Bibliotecas Digitais possam interoperar surgiu a *Open Archives Initiative* (OAI) (OAI, 2005) e, para resolver a questão da padronização dos metadados utilizados pelos repositórios, foi criado o formato Dublin Core (DUBLIN, 2005).

Neste capítulo, será apresentado o padrão OAI, incluindo o protocolo OAI-PMH e suas funcionalidades. Além disso, o formato Dublin Core será apresentado com as definições de seus elementos. Por fim, a BDBComp, uma biblioteca digital brasileira que utiliza tais tecnologias, será brevemente discutida, isto porque esta é utilizada como fonte de dados na avaliação experimental apresentada neste trabalho (capítulo 5).

2.1 *Open Archives Initiative* (OAI)

A *Open Archives Initiative* (OAI) surgiu com a Convenção de Santa Fé (*Santa Fe Convention*), realizada em Santa Fé, capital do Estado americano de Novo México, em 21-22 de outubro de 1999. Esta convenção apresenta um modelo técnico e organizacional simples para suportar interoperabilidade básica entre arquivos de *e-prints*. Após este encontro, a OAI passou por uma fase de desenvolvimento, sendo seus objetivos ampliados. Dessa forma, passou a ser aplicada a provedores de diversos tipos de conteúdo, principalmente de publicações científicas (SOMPEL; LAGOZE, 2000).

A *Open Archives Initiative* (OAI, 2005) teve um papel muito importante para permitir a interoperabilidade entre as Bibliotecas Digitais. Seu principal objetivo foi o de fazer com que diferentes Bibliotecas Digitais ao redor do mundo pudessem interoperar formando uma federação (SOMPEL; LAGOZE, 2000). Cabe salientar o sentido do nome *Open Archives Initiative* (Iniciativa de Arquivos Abertos), no qual, o

termo “arquivo” refere-se a repositórios para armazenamento de informações e o termo “aberto” refere-se à arquitetura do sistema que define interfaces, para facilitar a disponibilização de conteúdos de diferentes provedores.

2.1.1 OAI-PMH (*Open Archives Initiative - Protocol for Metadata Harvesting*)

A partir de então, ficou definida uma forma padrão de comunicação entre Bibliotecas Digitais. A OAI define o protocolo OAI-PMH (*Open Archives Initiative - Protocol for Metadata Harvesting*) (OAI-PMH, 2005), que provê um modelo para garantir interoperabilidade independente da aplicação baseado em colheita (*harvesting*) de metadados. Assim, a colheita de metadados por parte das Bibliotecas Digitais é feita através da utilização do protocolo OAI-PMH, que define como deve ser realizada a transferência de metadados entre duas entidades básicas: provedores de dados e provedores de serviços.

Os **provedores de dados** administram sistemas que suportam o protocolo OAI-PMH como meio de exposição de metadados. Um provedor de dados utiliza um repositório para exposição dos metadados para *harvesters*. Um **repositório** é um servidor de rede acessível que pode processar as seis requisições (verbos) do OAI-PMH (descritas nas seções seguintes).

Os **provedores de serviços** utilizam os metadados, obtidos através de colheita via OAI-PMH, como base para fornecer serviços mais específicos. Os provedores de serviços operam um ***harvester***, uma aplicação cliente que envia requisições OAI-PMH, como meio de realizar colheita de metadados dos repositórios.

O OAI-PMH faz distinção entre três entidades utilizadas para fazer os metadados acessíveis pelo OAI-PMH que são: **recurso**, **item** e **registro**.

- **Recurso** é um objeto que os metadados descrevem. A natureza do recurso, se físico ou digital, se é armazenado no repositório ou é constituinte de outra base de dados, está fora do escopo do OAI-PMH.
- **Item** é constituinte de um repositório através do qual metadados sobre o recurso podem ser disseminados. Um item é, conceitualmente, um “armazenador” de conteúdo que armazena ou dinamicamente gera metadados sobre um único recurso em múltiplos formatos, cada um podendo ser obtido como um registro via OAI-PMH. Cada item tem um identificador único no repositório do qual o item é constituinte. Um **identificador único**, como o próprio nome sugere, identifica um item, de forma não ambígua, dentro do repositório; o identificador único é usado nas requisições OAI-PMH para extração de metadados de um item. Itens podem conter metadados em múltiplos formatos. O identificador único mapeia o item, e todos os possíveis registros disponibilizados de um único item, de forma a compartilharem o mesmo identificador único.
- **Registro** são os metadados em um formato específico. Um registro é retornado, codificado em XML, em resposta a uma requisição do protocolo. Um registro é identificado de forma não ambígua pela combinação do identificador único do item do qual o registro é disponibilizado, o *metadataPrefix* identificando o formato de metadados do registro, e um rótulo de tempo (*datestamp*) (ex.: data de criação, modificação ou deleção do registro).

A interação entre as duas entidades básicas do OAI-PMH pode ser vista na Figura 2.1. Pode-se observar que um provedor de serviços que deseja realizar uma colheita de

metadados envia requisições HTTP para um provedor de dados que, de acordo com a requisição solicitada, envia como resposta os metadados solicitados em formato XML. Com base nos metadados recebidos, o provedor de serviços pode, então, oferecer um determinado serviço como, por exemplo, um sistema de busca ou recomendação.

Figura 2.1: Interação entre as entidades básicas do OAI-PMH (adaptado de OAI, 2005)

Para que seja possível a tarefa de colheita dos metadados de provedores de dados são definidos seis tipos de requisições chamadas de "verbos" que são: *Identify*, *ListMetadataFormats*, *ListSets*, *ListIdentifiers*, *ListRecords*, e *GetRecords*. Uma descrição sobre cada um desses verbos, incluindo os possíveis argumentos a serem utilizados, é apresentada nas seções seguintes. Para maiores detalhes ver (OAI-PMH, 2005).

As requisições do OAI-PMH são expressas como requisições HTTP (métodos GET ou POST). Existe uma URL base, para todas as requisições, que especifica o *host* do servidor HTTP atuando como um repositório. Em adição a URL base, todas as requisições consistem de uma lista de argumentos, dados na forma de pares *key=value*. Argumentos podem aparecer em qualquer ordem e múltiplos argumentos são separados por "&". Cada requisição OAI-PMH deve ter pelo menos um par *key=value* que especifica a requisição OAI-PMH enviada pelo *harvester*: onde *key* é a string "verb" e *value* é um dos seis verbos OAI-PMH definidos. O número e a natureza dos pares *key=value* adicionais dependem dos argumentos de cada requisição individual.

O OAI-PMH suporta **colheita seletiva**, ou seja, permite aos *harvesters* limitarem as requisições a porções dos metadados disponibilizados por um repositório, sendo possível a utilização de dois tipos de critérios que podem ser combinados nas requisições: rótulos de tempo (*datestamps*) e conjuntos (*sets*). A organização em conjuntos é uma construção opcional para agrupamento de itens. Repositórios podem organizar seus itens em conjuntos, sendo que esta organização pode ser simples ou hierárquica, incluindo a possibilidade de múltiplas hierarquias. A organização em conjuntos é expressa na sintaxe do parâmetro *setSpec*.

As respostas de todas as requisições do protocolo OAI-PMH são codificadas em XML, sendo que cada resposta inclui sua respectiva requisição. Além disso, o XML de cada resposta é definido segundo um XML Schema. Conforme Lagoze & Sompel (2001), isto objetiva a possibilidade de verificação da concordância com as especificações técnicas exigidas pelo OAI, permitindo que um programa de teste seja capaz de visitar um repositório OAI, enviar cada requisição do protocolo com vários argumentos e testar cada resposta conforme o esquema definido no protocolo para a resposta. Um exemplo de ferramenta padrão utilizada para realizar este tipo de teste é o *OAI Repository Explorer* (OAI Repository Explorer, 2006).

A seguir serão apresentados os seis verbos definidos pelo protocolo OAI-PMH.

2.1.1.1 Identify

O verbo *Identify* é usado para retornar informações sobre o repositório, tais como: nome, identificador, e-mail do administrador, informações sobre a propriedade intelectual dos dados contidos no repositório, etc. Nesta requisição nenhum argumento é requerido.

2.1.1.2 ListMetadataFormats

O verbo *ListMetadataFormats* é usado para obter informações sobre os formatos de metadados disponibilizados pelo repositório. Nesta requisição, há a possibilidade do uso do seguinte argumento para restringir a mesma aos formatos disponibilizados para descrever um item específico:

- ***identifier:*** especifica o identificador único do registro requerido.

2.1.1.3 ListSets

O verbo *ListSets* é usado para obter a estrutura de conjuntos de um repositório. O uso de conjuntos é uma possibilidade de organização dos registros oferecida pelo OAI-PMH. E, estando os registros classificados em conjuntos, a colheita seletiva de informações é facilitada. Nesta requisição o seguinte argumento pode ser requerido:

- ***resumptionToken:*** argumento exclusivo com um valor que é um *token* retornado para controle de fluxo por um *ListSets* anterior, que retornou uma lista incompleta.

2.1.1.4 ListIdentifiers

O verbo *ListIdentifiers* é uma abreviação do *ListRecords*, que obtém somente os identificadores de registros do repositório. Nesta requisição o seguinte argumento pode ser requerido:

- ***resumptionToken:*** argumento exclusivo com um valor que é um *token* retornado para controle de fluxo por um *ListIdentifiers* anterior, que retornou uma lista incompleta.

Além disso, exceto quando o argumento *resumptionToken* é usado, a requisição exige o seguinte argumento:

- ***metadataPrefix:*** especifica o *metadataPrefix* (identificador) do formato de metadados, em que os registros retornados devem estar descritos.

Argumentos opcionais permitem colheita seletiva de registros, baseada nos membros de um conjunto e/ou intervalos de datas:

- ***from:*** limite inferior para determinar o intervalo de datas base.
- ***until:*** limite superior para determinar o intervalo de datas base.
- ***set:*** valor que especifica um conjunto como critério de seleção.

2.1.1.5 ListRecords

O verbo *ListRecords* é usado para efetuar a colheita de registros de um repositório. Nesta requisição o seguinte argumento pode ser requerido:

- ***resumptionToken:*** argumento exclusivo com um valor que é um *token* retornado para controle de fluxo por um *ListRecords* anterior, que retornou uma lista incompleta.

Além disso, exceto quando o argumento *resumptionToken* é usado, a requisição exige o seguinte argumento:

- ***metadataPrefix:*** especifica o *metadataPrefix* do formato de metadados, em que os registros retornados devem estar descritos.

Alguns argumentos opcionais permitem colheita seletiva de registros, baseada nos membros de um conjunto e/ou intervalos de datas. Para tanto, os seguintes argumentos podem ser utilizados:

- ***from:*** limite inferior para determinar o intervalo de datas base.
- ***until:*** limite superior para determinar o intervalo de datas base.
- ***set:*** valor que especifica um conjunto como critério de seleção.

2.1.1.6 GetRecords

O verbo *GetRecords* é usado para obter os metadados de um registro individual de um repositório. Nesta requisição, os seguintes argumentos são requeridos:

- ***identifier:*** especifica o identificador único do registro requerido.
- ***metadataPrefix:*** indica o formato dos metadados a serem retornados.

2.2 *Dublin Core Metadata Initiative* (DCMI)

Outro ponto importante, para garantir interoperabilidade, é a adoção de um padrão básico para descrição dos metadados pelos provedores de dados que seguem o padrão OAI. Para tanto, foi escolhido o formato Dublin Core Simples, que permite, por sua simplicidade, a descrição dos recursos disponíveis na Internet, através de um conjunto mínimo de metadados, possuindo, ainda, um escopo internacional e podendo ser codificado em XML.

A definição do formato Dublin Core (DUBLIN, 2005) foi resultado do *OCLC/NCSA Metadata Workshop*, ocorrido em Dublin, Ohio, em março de 1995, que teve como discussão a semântica de metadados. Segundo (DCMI, 2006), neste evento, mais de 50 pessoas discutiram de que modo um conjunto núcleo, para descrever a semântica de recursos baseados na Web, poderia ser extremamente útil para categorizar a Web de forma a facilitar a busca e recuperação de informações.

O padrão Dublin Core inclui dois níveis: Simples e Qualificado.

O formato Dublin Core Simples é composto de quinze elementos, sendo que cada elemento é opcional e pode ser repetido para descrição de um dado recurso. O conjunto desses elementos (*Dublin Core Metadata Element Set* - DCMES) é apresentado sumarizado na Tabela 2.1, a descrição completa pode ser obtida em (DUBLIN, 2005). O conteúdo de alguns desses elementos pode ser determinado por um "vocabulário controlado", que é um conjunto limitado de termos bem definidos e consistentes (HILLMAN, 2005).

O Dublin Core Qualificado inclui três elementos adicionais, apresentados na Tabela 2.2, bem como um grupo de elementos de refinamento (também chamados de qualificadores) que refina a semântica dos elementos, de modo a ser útil na descoberta de recursos. A semântica do Dublin Core tem sido estabelecida por um grupo internacional e multi-disciplinar de profissionais (HILLMAN, 2005).

Tabela 2.1: Conjunto de elementos do formato Dublin Core

Nome do Elemento	*Definição*
dc:title	Um nome dado ao recurso. (ex.: título)
dc:creator	Uma entidade primariamente responsável pela criação do conteúdo do recurso. (ex.: autores)
dc:subject	Um tópico de conteúdo do recurso. (ex.: palavras-chave)
dc:description	Um apanhado do conteúdo do recurso. (ex.: resumo/abstract)
dc:publisher	Uma entidade responsável pela disponibilização do recurso. (ex.: editora)
dc:contributor	Uma entidade responsável por fazer contribuições ao conteúdo do recurso.
dc:date	Uma data de um evento do ciclo de vida do recurso. (tipicamente, *dc:date* será associada com a criação ou disponibilização do recurso)
dc:type	A natureza ou gênero do conteúdo do recurso.
dc:format	A manifestação física ou digital do recurso.
dc:identifier	Uma referência, não ambígua do recurso, dentro de um dado contexto. (ex.: URL, ISBN)
dc:source	Uma referência para um recurso do qual o presente recurso é derivado.
dc:language	Idioma no qual o conteúdo intelectual do recurso está escrito.
dc:relation	Uma referência para um recurso relacionado.
dc:coverage	A extensão ou escopo do conteúdo do recurso. (tipicamente, *dc:coverage* irá incluir uma localização geográfica)
dc:rights	Informações sobre direitos, propriedade intelectual ou condições de uso do recurso.

Tabela 2.2: Conjunto de elementos adicionais do formato Dublin Core Qualificado

Nome do Elemento	*Definição*
dc_qual:audience	Uma entidade para a qual o recurso é dirigido ou útil.
dc_qual:provenance	Indicação de mudanças na posse e/ou custódia do recurso, desde a sua criação, que forem significativas para garantir sua autenticidade, integridade e interpretação.
dc_qual:rightsHolder	Uma pessoa ou organização que possui ou controla os direitos sobre o recurso.

2.3 Biblioteca Digital Brasileira de Computação (BDBComp)

Uma iniciativa nacional em bibliotecas digitais é a BDBComp (Biblioteca Digital Brasileira de Computação) (BDBComp, 2005), projeto do grupo de Banco de Dados da Universidade Federal de Minas Gerais (UFMG).

A BDBComp tem como objetivo prover uma plataforma para arquivamento, indexação, disseminação e preservação do conhecimento científico produzido pela comunidade brasileira da área da Ciência da Computação (SILVA; LAENDER; GONÇALVES, 2005). A BDBComp é desenvolvida em conformidade com o padrão OAI e adota o formato Dublin Core (DC) como padrão de metadados. Assim, é possível, via protocolo OAI-PMH, efetuar a colheita dos metadados presentes em tal repositório.

Os metadados disponibilizados pela BDBComp podem ser informados através do serviço de auto-arquivamento e ainda existem outros dois meios alternativos de obter-se metadados para o repositório da BDBComp: a extração desses de páginas Web e a colheita em outros repositórios que sigam o padrão OAI (CITIDEL por exemplo).

A seguir serão apresentados, na Tabela 2.3, os metadados disponibilizados pela BDBComp sobre os artigos científicos indexados pela mesma. À esquerda, temos os elementos do DC e, à direita, uma explicação mais específica sobre a sua utilização na BDBComp.

É importante observar que os metadados apresentados, indexados pela BDBComp, servirão de base para o sistema de recomendação proposto neste trabalho. Os mesmos descreverão os artigos que possivelmente serão recomendados pelo sistema.

Tabela 2.3: Elementos do DC utilizados pela BDBComp

Elementos DC	*BDBComp*
dc:title	Título do artigo.
dc:creator	Autor do artigo. (elemento repetido de acordo com o número de autores)
dc:subject	Palavras-chave. (este metadado não está disponível na descrição de nenhum dos artigos indexados pela BDBComp até junho de 2006)
dc:description	Resumo.
dc:publisher	Publicador dos anais (*proceedings*).
dc:contributor	Não utilizado.
dc:date	Ano de publicação.
dc:type	*Text* (termo do vocabulário controlado, que indica um recurso consistindo basicamente de texto para leitura)
dc:format	Formato do documento. (ex.: pdf)
dc:identifier	URL e identificador na BDBComp. (elemento é repetido para representar estas duas informações)
dc:source	Não utilizado.
dc:language	Idioma.
dc:relation	Não utilizado.
dc:coverage	Local do evento em que o artigo foi publicado.
dc:rights	Copyrights.

3 SISTEMAS DE RECOMENDAÇÃO

Sistemas de recomendação provêm uma interface alternativa para tecnologias de filtragem e recuperação de informações, tendo como foco a predição daqueles itens ou partes da informação que o usuário acharia interessante e útil. Tais predições são personalizadas, baseadas no perfil de cada usuário e podem conter julgamentos de interesses ou grau de relevância de itens previamente vistos pelo usuário. Segundo Herlocker (2000), um dos desafios consiste em coletar informações sobre as preferências dos usuários. Esta coleta de informações pode ser realizada de maneira explícita, solicitando ao usuário que especifique suas preferências, através do uso de coeficientes de avaliação numéricos para os itens. Porém convém salientar que, esta forma gera uma sobrecarga adicional aos usuários do sistema. Ou ainda, de maneira implícita, em que o sistema infere os coeficientes através da observação das ações do usuário com o sistema. Detalhes sobre o perfil do usuário utilizado no sistema de recomendação proposto são apresentados na seção 4.1 deste trabalho.

3.1 Abordagens para filtragem de informação

Para sistemas de recomendação, geralmente, três tipos de informações são disponibilizados: informação do item (descrição textual do conteúdo dos itens a serem recomendados), informação do usuário (que receberá a recomendação) e informação transacional (históricos dos itens recomendados aos usuários e as avaliações destes). As fontes de informações utilizadas no nível de representação determinam o tipo de abordagem adotada. Dessa forma, em relação aos sistemas de recomendação, sob o ponto de vista metodológico, existem três tipos básicos de abordagens utilizadas (HERLOCKER, 2000; HUANG et al., 2002; BALABANOVIC; SHOHAM, 1997; CLAYPOOL et al., 1999): filtragem baseada em conteúdo, filtragem colaborativa e filtragem híbrida, que serão comentadas nas seções seguintes. Uma abordagem de filtragem colaborativa, usualmente, lida com informações transacionais, enquanto abordagens baseadas em conteúdo utilizam informações do item e ambas, possivelmente, utilizam informações do usuário. Já em sistemas híbridos, há a tentativa de combinar essas três diferentes fontes de informação no processo de geração das recomendações.

3.1.1 Filtragem baseada em conteúdo (*Content-based Filtering*)

A abordagem de filtragem baseada em conteúdo possui este nome devido ao fato de os sistemas, que a adotam, desenvolverem a filtragem baseada em análises dos conteúdos dos itens, que possivelmente serão recomendados e podendo, também, utilizar informações do perfil do usuário. Esta abordagem trabalha com a idéia de gerar recomendações de itens relacionados ao perfil do usuário. Um perfil do item consiste de

alguns atributos, que descrevam o conteúdo do item, e o perfil do usuário é criado, com base em informações, que descrevam os interesses do usuário, e relacionadas com o perfil dos itens. A recomendação é gerada utilizando algumas funções de similaridade para fazer o casamento desses perfis (HUANG et al., 2002).

Para tanto, a informação precisa ser automaticamente reconhecida e categorizada, sendo gerados descritores do conteúdo de cada item. As descrições das necessidades de interesse do usuário são, ou supridas pelo usuário, como uma consulta, ou apreendidas pela observação do conteúdo dos itens consumidos pelo usuário (HERLOCKER, 2000). Então, a comparação da descrição de cada item com a descrição da necessidade de informação do usuário é utilizada para determinar se um item é ou não relevante para atender as necessidades do usuário.

O sistema de recomendação, apresentado neste trabalho, utiliza a abordagem baseada em conteúdo. Isto porque a idéia, neste caso, é gerar uma recomendação baseada no conteúdo, que combine as informações do usuário, obtidas a partir do seu currículo Lattes (principalmente referentes ao conteúdo de seus trabalhos desenvolvidos), com as informações referentes aos artigos de bibliotecas digitais para gerar a recomendação personalizada. Portanto, exemplos de tecnologias aplicadas para filtragens baseadas em conteúdo são discutidas na seção 3.2, sendo modelos clássicos utilizados para recuperação de informação, já que a abordagem baseada em conteúdo é derivada dos conceitos introduzidos pela comunidade de Recuperação de Informação (IR) (SHAHABI; CHEN, 2003).

A filtragem baseada em conteúdo possui algumas limitações como: o conteúdo de dados pouco estruturados é de difícil análise (por exemplo: imagens, vídeos e sons); o processamento do conteúdo do texto pode ser prejudicado devido ao uso de termos sinônimos; pode ocorrer a "super especialização", pois o sistema não recomenda itens cujo conteúdo não "case" com o perfil do usuário (CAZELLA; REATEGUI, 2005). Dessa maneira, neste tipo de abordagem, não existe "surpresa" na recomendação, já que itens que não se relacionam com o perfil do usuário não serão recomendados a este. Além disso, segundo Claypool et al. (1999), técnicas baseadas em conteúdo têm a dificuldade de distinguir entre informação de alta e de baixa qualidade sobre o mesmo tópico. Outra limitação acontece, caso o perfil do usuário seja construído a partir de informações obtidas pela interação deste com o sistema. Neste caso, há a necessidade do usuário ter avaliado um número suficiente de itens, antes que o sistema de recomendação possa realmente "entender" as preferências do usuário e apresentar recomendações confiáveis. Estas recomendações serão baseadas no "casamento" entre o conteúdo dos itens a serem recomendados e o conteúdo dos itens preferidos pelo usuário (ADOMAVICIUS; TUZHILIN, 2005).

3.1.2 Filtragem colaborativa (*Collaborative Filtering*)

Na filtragem colaborativa as ações do usuário e análises a respeito de uma informação particular são registradas para benefício de uma comunidade maior. Membros de uma comunidade podem beneficiar-se de experiências de outros, antes de consumir uma nova informação (HERLOCKER, 2000). Esta abordagem não requer nenhum tipo de descrição do conteúdo do item para que este seja recomendado. Por esta razão, a abordagem tem sido desenvolvida para cobrir áreas, onde a filtragem baseada em conteúdo é fraca.

A filtragem colaborativa utiliza a opinião de outros usuários a respeito da

informação a ser recomendada. Sistemas desse tipo podem ser não-personalizados, permitindo ao usuário descobrir itens que são de interesse popular e evitar os de desagrado popular, e podem ser personalizados, através dos relacionamentos entre perfis de usuários, trabalhando com a idéia de que, se os interesses dos usuários são similares, itens preferidos por um usuário podem ser recomendados a outros usuários com perfil similar (ou à comunidade que este usuário faz parte). Tais relacionamentos entre usuários podem ser informados ao sistema ou descobertos de forma automática, com base na análise de padrões comuns nas avaliações dos itens.

Nesse tipo de abordagem, podem ocorrer problemas como a "partida fria" (*coldstart*) quando não estão inicialmente disponíveis dados sobre o perfil do usuário, não havendo informações que possibilitem encontrar um perfil similar. Ou ainda, segundo Balabanovic & Shoham (1997), se um novo item for adicionado na base de dados, não existe meio deste ser recomendado até que um usuário o avalie ou especifique outro item já avaliado como similar a este. Quando há um número de usuários relativamente pequeno para o volume de informação do sistema, existe o risco de a cobertura das avaliações dos itens tornar-se muito esparsa, diminuindo a coleção de itens recomendáveis. Mais ainda, recomendações de itens recentes na base de dados podem ser inexatas, porque existem poucas avaliações para basear as predições da filtragem colaborativa. Além disso, segundo Claypool et al. (1999), em pequenas ou até médias comunidades de usuários, existem indivíduos que não se beneficiam de sistemas de filtragem colaborativa puros, porque suas opiniões não estão consistentemente de acordo ou em desacordo com qualquer grupo de pessoas (usuários com gostos incomuns).

3.1.3 Filtragem Híbrida (*Hybrid Filtering*)

As abordagens, de filtragem baseada em conteúdo e filtragem colaborativa, não são mutuamente exclusivas, existindo inúmeros esforços para integração de ambas, a fim de obter maior exatidão nas recomendações (HUANG et al., 2002). Para tanto, a abordagem de filtragem híbrida surge como uma combinação dessas duas abordagens (apresentadas anteriormente nas seções 3.1.1 e 3.1.2), buscando agregar as características de cada uma delas e solucionar as limitações encontradas, de forma a melhor atender as necessidades dos usuários. Segundo Claypool et al. (1999), usando esta combinação, podem ser alcançados os benefícios da filtragem baseada em conteúdo, que inclui a predição para todos os itens e usuários (sem a dependência do número de usuários e do número de avaliações dos itens), enquanto se ganha em exatidão nas predições de filtragem colaborativa conforme o número de usuários e avaliações cresce.

Segundo Cazella & Reategui (2005), algumas das características importantes, herdadas pela filtragem híbrida de cada uma das abordagens, podem ser observadas na Figura 3.1, são elas: (i) descoberta de novos relacionamentos entre usuários; (ii) recomendação de itens diretamente relacionados ao histórico; (iii) bons resultados para usuários incomuns; (iv) precisão independente do número de usuários.

As características (i) e (ii) são herdadas da filtragem colaborativa, já que esta trabalha com a idéia de "perfis similares". Na filtragem baseada em conteúdo não é levado em consideração qualquer tipo de relacionamento entre perfis de usuários. Além disso, itens com histórico de boa recepção por diversos tipos de usuários também não são relevantes na filtragem baseada em conteúdo, pois esta não gera a recomendação de itens não relacionados ao perfil do usuário (sem "surpresa" na recomendação).

As características (iii) e (iv) são alcançadas graças à abordagem baseada em conteúdo. Se fosse utilizada somente a filtragem colaborativa, não seria possível obter bons resultados para usuários incomuns, pois não se conseguiria um perfil de usuário semelhante para "casar" com o perfil destes usuários, assim como, havendo poucos usuários surge dificuldade na obtenção de informações para casamento entre os perfis. Isto já não ocorre na filtragem baseada em conteúdo.

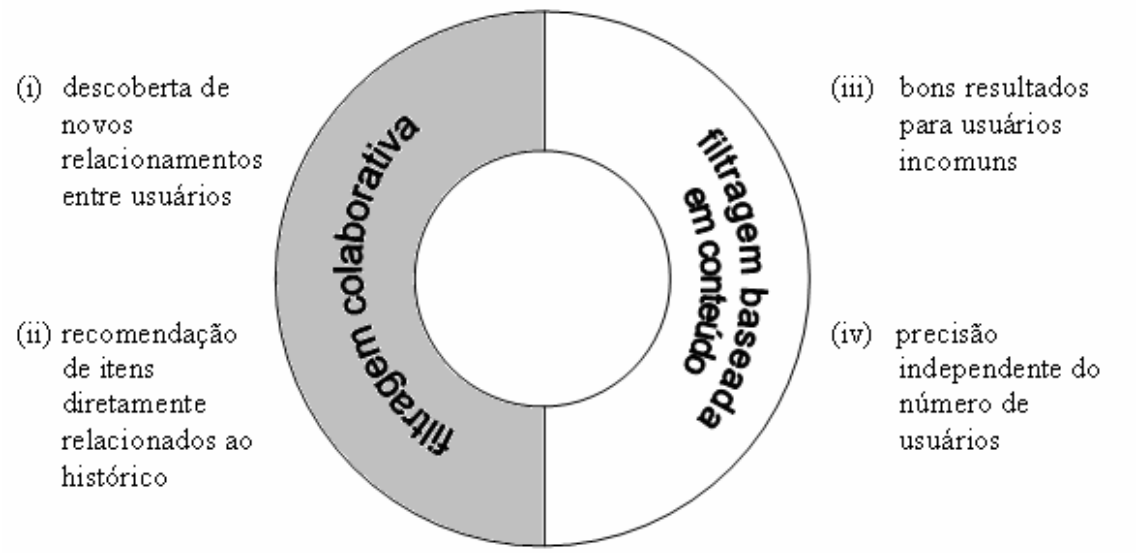

Figura 3.1: Características herdadas pela Filtragem Híbrida (adaptado de CAZELLA; REATEGUI, 2005)

Segundo Huang et al. (2002), sistemas híbridos podem obter diferentes graus de ganho em exatidão de predição, por utilizarem múltiplas fontes de informação (informações do item, do usuário e transacionais), variando de modestos benefícios a melhorias significativas. Porém, esta adição de informação nem sempre conduz a melhores resultados. A análise da variação na qualidade da recomendação, em função da multidimensionalidade da informação, requer um estudo muito aprofundado.

3.2 Modelos para Recuperação de Informação

O processo de recomendação pode ser visto como recuperação de informação, no qual documentos relevantes aos usuários devem ser recuperados e recomendados. A fim de gerar recomendações, podem ser utilizados os modelos clássicos de recuperação de informação tais como: modelo booleano, modelo de espaço vetorial e modelo probabilístico (SALTON; MCGILL, 1983; BAEZA-YATES; RIBEIRO-NETO, 1999; GROSSMAN, 2004). Tais modelos são apresentados nas seções seguintes.

Os modelos clássicos consideram cada documento sendo descrito por um conjunto de termos de indexação, que são palavras ou expressões usadas para identificação e representação do conteúdo do documento. Esses modelos não levam em consideração a correlação entre os termos de indexação o que é claramente uma simplificação adotada, que pode, teoricamente, constituir uma desvantagem. Entretanto, na prática, a consideração da dependência dos termos pode acabar constituindo uma desvantagem, já que sua aplicação indiscriminada em todos os documentos da coleção pode acabar diminuindo a performance total do sistema de recuperação. Assim, não é claro que a

suposição da independência dos termos de indexação seja ruim em situações práticas (BAEZA-YATES; RIBEIRO-NETO, 1999).

3.2.1 Modelo Booleano

O modelo booleano (SALTON; MCGILL, 1983) é um modelo de recuperação de informação simples, baseado em teoria de conjuntos e álgebra booleana. Neste modelo, os documentos (D) são representados como conjuntos de termos de indexação e as consultas (Q) são formuladas através de expressões booleanas formadas por termos e conectivos de *boole* (*and*, *or* e *not*). Estas expressões booleanas permitem a combinação das operações de união (*or*), intersecção (*and*) e negação (*not*) de conjuntos. Um exemplo de uma expressão booleana que poderia ter sido formulada, visando recuperar documentos que possuam informações sobre *sistemas* de *recomendação* e/ou sobre *OAI*, é apresentado na Figura 3.2.

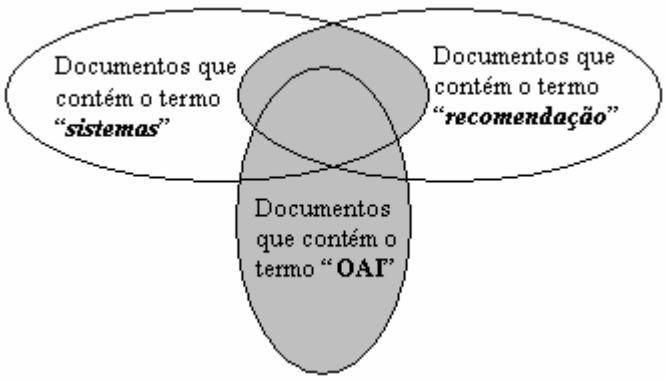

Figura 3.2: Representação gráfica do resultado da expressão booleana ***("sistemas" and "recomendação") or "OAI"***

Como resultado do processo de recuperação, são retornados somente os documentos que satisfazem todas as restrições lógicas representadas pela consulta, ou seja, é uma estratégia de recuperação, baseada num critério de decisão binário (apresentado na Equação 3.1). Um documento só pode ser dito como relevante (*Similaridade(Q, D) = 1*) ou não-relevante (*Similaridade(Q, D)* = 0). Neste modelo não há a noção de "casamento parcial" em relação às condições da consulta.

$$Similaridade(Q, D) = \begin{cases} 1 & \text{se } D \text{ satisfaz condições da expressão booleana } Q \\ 0 & \text{caso contrário} \end{cases} \quad (3.1)$$

Para garantir a qualidade da informação recuperada, muitas vezes, há a necessidade de especificações de consultas complexas, o que exige um conhecimento profundo da informação desejada e sobre lógica booleana na elaboração de consultas. Outra dificuldade do uso do modelo booleano é a incapacidade de se representar pesos associados aos termos desejados, não havendo possibilidade de diferenciação entre a importância dos termos para consulta desejada. Ou seja, dois termos, solicitados em uma consulta, são igualmente importantes para representar a mesma, e, de forma análoga, se um termo está presente em um documento ele é tão importante para representação daquele documento quanto qualquer outro termo presente no mesmo. Portanto, o modelo booleano não leva em consideração as diferentes importâncias dos termos para representação dos documentos.

Segundo Baeza-Yates & Ribeiro-Neto (1999), as principais vantagens do modelo booleano são sua simplicidade e o formalismo claro envolvido; e as desvantagens estão associadas ao fato de que o "casamento" exato pode levar à recuperação de poucos ou muitos documentos, obtendo uma performance fraca, que o leva a ser considerado o mais fraco dos modelos clássicos.

Para resolver algumas limitações do modelo booleano, foi proposta uma extensão a este: o Modelo Booleano Estendido (SALTON; FOX; WU, 1983), um modelo intermediário entre o sistema booleano e o modelo vetorial (apresentado na seção 3.2.2). A estrutura de consulta inerente ao sistema booleano é preservada, e, ao mesmo tempo, pesos associados aos termos podem ser incorporados em consultas e documentos. A saída obtida pode então ser classificada em ordem de similaridade com a consulta desejada. Tal modelo torna-se ainda mais complexo porque, além da necessidade do domínio da lógica booleana, existe a necessidade da determinação do grau de importância dos termos representado através dos valores de seus pesos (tanto para termos da consulta quanto para representação dos documentos).

3.2.2 Modelo de Espaço Vetorial (VSM) ou Modelo Vetorial

Neste trabalho, o VSM (*Vector Space Model*) é selecionado por ser um modelo adequado às necessidades do trabalho proposto: um modelo baseado em conteúdo, com pesos associados aos termos de indexação e cujo resultado da função de similaridade é dado na forma de *ranking*. Além disso, tal abordagem possui uma implementação relativamente simples e provê resultados satisfatórios. No modelo de espaço vetorial (SALTON; WONG; YANG, 1975), documentos e consultas são representados como vetores de termos de indexação. Cada termo tem um peso associado a si, para prover distinção entre os termos de acordo com sua importância. Segundo Salton & Buckley (1988), os pesos podem variar entre 0 e 1. Valores próximos a 1 correspondem a termos mais importantes, enquanto valores próximos a 0 correspondem a termos menos relevantes.

O VSM utiliza um espaço *n*-dimensional para representar os termos, onde *n* corresponde ao número de termos distintos. Para cada vetor de documentos, ou de consulta, os pesos representam as coordenadas do vetor na dimensão correspondente. O princípio do VSM é baseado na correlação inversa entre a distância (ângulo) entre vetores de termos no espaço e a similaridade entre os documentos que eles representam.

Para calcular o escore de similaridade, o co-seno (Equação 3.2) pode ser utilizado (fórmula obtida do produto escalar entre os dois vetores desejados, dividido pela multiplicação dos módulos desses vetores). O valor resultante indica o grau de relevância entre a consulta (Q) e o documento (D), onde w representa os pesos dos termos contidos em Q e D, e t representa o número de termos (tamanho do vetor). Esta equação provê uma saída, classificada com base na ordem decrescente dos valores de similaridade obtidos (SALTON; BUCKLEY, 1988).

$$Similaridade(Q, D) = \frac{\sum_{k=1}^{t} w_{qk} \cdot w_{dk}}{\sqrt{\sum_{k=1}^{t} (w_{qk})^2 \cdot \sum_{k=1}^{t} (w_{dk})^2}} \tag{3.2}$$

Na Figura 3.3, é apresentado um exemplo de um espaço vetorial tri-dimensional, onde cada dimensão é um dos termos de indexação, no caso t_1, t_2 e t_3. A figura mostra dois documentos, D_1 (0,5; 0,0; 0,45) e D_2 (0,4; 0,55; 0,05), representados neste espaço vetorial. Os números indicam as coordenadas dos termos nessas dimensões e representam os pesos associados a cada um dos termos na representação dos documentos. Além disso, é representada, neste espaço vetorial, a consulta Q (0,3; 0,5; 0,1) desejada, onde o peso dos termos representa a respectiva importância de cada termo para esta consulta. Sendo aplicada a Equação 3.2, para calcular a similaridade entre os documentos e a consulta desejada, são obtidos os seguintes valores aproximados:

- *Similaridade* $(Q, D_1) = 0{,}4899 = 48{,}99\%$
- *Similaridade* $(Q, D_2) = 0{,}9915 = 99{,}15\%$

Com este resultado, pode ser observado que o documento D_2 é mais similar à consulta Q, do que o documento D_1, conforme evidenciado na representação gráfica da Figura 3.3, onde D_2 está mais próximo à consulta Q (ângulo formado entre os dois vetores é menor) do que D_1.

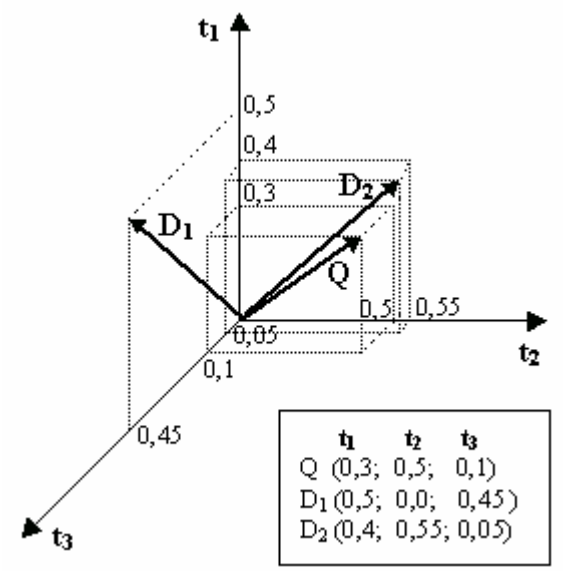

Figura 3.3: Representação de um espaço vetorial tri-dimensional

A Equação 3.2 é amplamente utilizada para comparar a similaridade entre documentos, e similarmente, em nosso caso, Q representa o perfil do usuário e D os descritores dos documentos que foram obtidos através de colheita (*harvesting*) em uma biblioteca digital (ver capítulo 4 para detalhes). O sistema de atribuição de pesos aos termos é muito importante para garantir um processo de recuperação de informação efetivo. Os resultados deste processo dependem crucialmente da escolha deste sistema. Além disso, a seleção dos termos da consulta é fundamental para a obtenção de um resultado de recomendação, em conformidade com as necessidades do usuário. Com relação a estas questões, serão apresentadas, mais adiante (na seção 4.3), as decisões adotadas para o sistema de recomendação proposto neste trabalho.

Segundo Baeza-Yates & Ribeiro-Neto (1999), as principais vantagens do modelo vetorial são: (i) esquema de atribuição de pesos aos termos, melhora a performance da recuperação; (ii) estratégia de casamento parcial, que permite recuperação de documentos, os quais se "aproximam" das condições da consulta; (iii) fórmula de *ranking* pelo co-seno ordena os documentos, de acordo com o grau de similaridade, em

relação à consulta. Conceitualmente, o modelo vetorial tem a desvantagem de não considerar a correlação entre os termos de indexação (discutida no início da seção 3.2).

3.2.3 Modelo Probabilístico

O modelo probabilístico, como o próprio nome sugere, utiliza a teoria das probabilidades como meio para modelar o processo de recuperação de informação. Neste modelo, a função de similaridade de um documento, para responder a uma expressão de busca, é calculada pela probabilidade de um documento (D) ser relevante a uma consulta (Q), caso os termos (t_i), especificados nesta consulta, apareçam no documento. Presume-se que a distribuição dos termos, nos documentos da coleção, é uma informação capaz de determinar a relevância ou não de um documento em responder a uma dada consulta. Portanto, a idéia de tal modelo é de que, quando vetores de documentos e consultas (termos envolvidos em ambas) são suficientemente similares, a probabilidade de relevância correspondente é alta o suficiente, para ser razoável recuperar o documento em resposta à consulta (SALTON; MCGILL, 1983).

Sendo assim, cada documento é representado por um vetor de termos, só que, diferentemente do modelo de espaço vetorial (seção 3.2.2), não há um peso associado a cada termo e sim um valor binário associado, que apenas indica a presença (1), ou ausência (0) do termo no documento. A função de similaridade é calculada utilizando-se a Equação 3.3, onde: $p(t_i|Rel)$ é a probabilidade de um termo t_i estar presente em um documento selecionado do conjunto dos relevantes, $p(t_i|\overline{Rel})$ é a probabilidade de um termo t_i estar presente em um documento selecionado do conjunto dos não-relevantes, $p(\overline{t_i}|\overline{Rel})$ é a probabilidade de um termo t_i não estar em um documento selecionado do conjunto dos não-relevantes e $p(\overline{t_i}|Rel)$ é a probabilidade de um termo t_i não estar presente em um documento selecionado do conjunto dos documentos relevantes.

$$Similaridade(Q,D) = \sum_{i=1}^{t}\left(\log \frac{p(t_i \mid \mathrm{Re}\, l).p(\overline{t_i} \mid \overline{\mathrm{Re}\, l})}{p(t_i \mid \overline{\mathrm{Re}\, l}).p(\overline{t_i} \mid \mathrm{Re}\, l)}\right) \tag{3.3}$$

A Equação 3.3 é fundamental para ordenar os documentos no modelo probabilístico, sendo obtida com base na aplicação do Teorema de Bayes. Esta equação pode ser expressa na forma da Equação 3.4, se forem considerados os parâmetros apresentados na Tabela 3.1, onde: N representa o número total de documentos da coleção; n é o número de documentos, contendo o termo desejado; R é o número de documentos (D), relevantes para a consulta (Q) e r é o número desses documentos relevantes que contêm o termo especificado (para simplificar, o sufixo i, que especifica o termo desejado, foi omitido; $r = r_i$ e $n = n_i$ são para termos especificados).

$$Similaridade(Q,D) = \sum_{i=1}^{t}\left(\log \frac{r.(N-n-R+r)}{(n-r).(R-r)}\right) \tag{3.4}$$

Tabela 3.1: Tabela de contingência da incidência de termos

	Relevantes	*Não-Relevantes*	
Contendo termo	r	n – r	n
Não contendo termo	R – r	N – n – R + r	N – n
	R	N – R	N

Fonte: Adaptado de (JONES; WALKER; ROBERTSON, 1998).

Segundo Baeza-Yates & Ribeiro-Neto (1999), a principal vantagem do modelo probabilístico, em teoria, é o *ranking* dos documentos ser realizado em ordem decrescente da probabilidade de relevância e as desvantagens são: (i) necessidade da suposição da separação inicial dos documentos nos conjuntos relevantes ou não-relevantes; (ii) o fato do método não levar em conta a freqüência com que um termo de indexação ocorre no documento (pesos binários); (iii) a adoção da suposição da independência dos termos (discutida no início da seção 3.1). Além disso, há controvérsias sobre o melhor desempenho do modelo probabilístico, em relação ao modelo vetorial, para maiores detalhes ver (BAEZA-YATES; RIBEIRO-NETO, 1999).

3.3 Exemplos de Sistemas de Recomendação para Bibliotecas Digitais

A seguir, serão apresentados alguns Sistemas de Recomendação, propostos no contexto de Bibliotecas Digitais, sendo feita uma breve apresentação de cada um deles, discutidos os propósitos da recomendação a ser realizada e analisado o tipo de abordagem para filtragem de informação adotado.

3.3.1 Sistema de recomendação baseado em grafo

O trabalho de (HUANG et al., 2002) apresenta um modelo de grafo de duas camadas, no contexto da recomendação de livros, este modelo é genérico, possibilitando a utilização das três abordagens: filtragem baseada em conteúdo, filtragem colaborativa e filtragem híbrida. Os testes do sistema desenvolvido foram realizados com informações de uma livraria chinesa *online* de Taiwan. O sistema utiliza informações sobre o conteúdo dos livros (*book*), informações demográficas sobre os clientes (*customer*), e seus históricos de compras (respectivamente similares, em Bibliotecas Digitais, à informação do conteúdo dos documentos, atributos pessoais dos usuários e seus históricos de uso). O método proposto utiliza um grafo de duas camadas (*book layer*, *customer layer*) e ainda, ligações entre as duas camadas, que representam o histórico de compra relacionando usuários a itens (*purchase history*).

Segundo Huang et al. (2002), o tipo de abordagem considerado irá variar de acordo com os pesos de similaridade considerados para predição dos itens a serem recomendados. Se forem utilizados somente os pesos da similaridade *book-to-book* (camada *book layer*), tem-se uma abordagem puramente baseada em conteúdo. Se forem utilizados somente os pesos da similaridade *customer-to-customer* (camada *customer layer*) e históricos de compras (*purchase histories*), para gerar a recomendação, tem-se uma abordagem puramente de filtragem colaborativa. Também pode-se combinar ambas as abordagens, pelo uso de todos os pesos de associação e histórico de compras, sendo considerada uma abordagem híbrida.

3.3.2 Sistema de recomendação de literatura

O sistema de recomendação proposto em (HWANG; HSIUNG; YANG, 2003) está inserido no projeto *Networked Digital Library Project* da *National Sun Yat-sem University* em Taiwan. O principal objetivo deste projeto é o desenvolvimento de tecnologias para suportar serviços digitais, sendo, uma das etapas, o desenvolvimento de um sistema de recomendação de literatura.

Esse sistema de recomendação de literatura emprega logs de uso Web da literatura da biblioteca digital. O modelo deste consiste de três passos seqüênciais: preparação dos dados dos logs de uso Web (mineração de uso na Web), descoberta da associação dos artigos, e recomendação de artigos. A recomendação busca predizer artigos relevantes para pesquisadores. O sistema possui uma interface com o usuário, sendo que o núcleo é um mecanismo recomendador, que analisa o uso da literatura, gerando recomendações classificadas de acordo com as preferências do usuário ativo no sistema. Várias características das publicações e das interações WWW são levadas em consideração. Um perfil de tarefas é utilizado para o usuário (conjunto de itens recentemente acessados), em vez de um perfil de interesses de longo-prazo.

A abordagem utilizada é conhecida como focada em tarefa (*task-focused approach*), esta representa uma combinação das idéias de filtragem colaborativa e mineração de dados (*data mining*) (HERLOCKER; KONSTAN, 2001). No sistema em questão, são montados *clusters* (aglomerados) de artigos acessados, freqüentemente juntos, que irão servir de base para geração da recomendação, de acordo com o perfil do usuário (artigos acessados por ele). Portanto, esta abordagem pode ser considerada uma variação da abordagem de filtragem colaborativa, que utiliza um perfil de usuário de acordo com suas interações recentes no sistema (mineração de dados) e faz o "casamento" entre este perfil e *clusters*, construídos com base no comportamento de acesso realizado por outros usuários; sem, no entanto, haver um "casamento" explícito entre perfis de usuários. Além disso, esta recomendação é feita independente do conteúdo dos itens a serem recomendados.

3.3.3 Sistema colaborativo personalizado

O trabalho apresentado em (RENDA; STRACCIA, 2002) foi desenvolvido no contexto do projeto CYCLADES, cujo objetivo é o provimento de um ambiente integrado de usuários e grupos de usuários (comunidades) que desejam usar, de forma personalizada e flexível, documentos digitais, dentro do contexto de bibliotecas digitais que sigam o padrão OAI. Dentre os serviços oferecidos pelo sistema, é disponibilizado um serviço de recomendação de documentos.

O ambiente desenvolvido permite que os usuários organizem o seu espaço de informação (diretórios), de acordo com seus próprios pontos de vista, suportando um ambiente colaborativo, onde é possível prover funções de recomendação. No modelo de recomendação proposto em (RENDA; STRACCIA, 2002) são utilizadas, em conjunto, as abordagens baseada em conteúdo e colaborativa, para computar o valor de similaridade entre os diretórios (abordagem híbrida), sendo possível a recomendação de itens de dados pertencentes a outros diretórios similares aos do usuário (documentos estes que devem ser distintos dos já pertencentes à hierarquia de diretórios do usuário).

3.3.4 *TalkMine*

O sistema de recomendação *TalkMine* foi desenvovido para a biblioteca de pesquisa do *Los Alamos National Laboratory*, sendo parte integrante do projeto *Adaptative Recommendation Project* (ARP, 2006). Este projeto visa a pesquisa e o desenvolvimento de sistemas de recomendação para bibliotecas digitais.

O sistema possui uma interface de busca a documentos da biblioteca digital. Dependendo das pesquisas realizadas pelo usuário, são definidas diferentes "personalidades de busca" para este, com históricos de IR (palavras-chave utilizadas na busca) diferentes e contextos de conhecimento independentes. Dessa forma, o algoritmo de recomendação integra o contexto de conhecimento da personalidade corrente do usuário, com os recursos de informação (documentos) buscados, possibilitando recomendações apropriadas. Além disso, o comportamento de todos os usuários do sistema é utilizado para adaptar os contextos de conhecimento dos recursos de informação pesquisados. Esta adaptação permite aos recursos de informação responderem melhor às expectativas dos usuários. Segundo Rocha (2001), o sistema é implementado com ambas as abordagens: baseada em conteúdo e colaborativa, caracterizando-se como uma abordagem híbrida.

4 SISTEMA DE RECOMENDAÇÃO PROPOSTO

O Sistema de Recomendação proposto tem como foco a recomendação de artigos científicos para a comunidade da área da Ciência da Computação. O perfil do usuário utilizado é um subconjunto do currículo Lattes. Uma forma alternativa ou complementar para geração deste perfil pode ser desenvolvida, como, por exemplo: um sistema de recuperação de informação para coleta de dados de páginas pessoais e/ou outras fontes de dados, a fim de gerar um arquivo XML, equivalente ao mesmo subconjunto do Lattes utilizado neste sistema de recomendação. Isto possibilita que usuários, não possuidores do currículo Lattes, como pesquisadores estrangeiros, também possam utilizar o sistema de recomendação proposto. Qualquer biblioteca digital, que provê seus metadados no formato Dublin Core (DC) e suporta o protocolo OAI-PMH, pode ser utilizada como fonte para prover informações sobre os artigos a serem recomendados. Na avaliação experimental, apresentada no capítulo 5, é utilizada a Biblioteca Digital Brasileira de Computação (BDBComp) (BDBComp, 2005).

Um provedor de dados de uma biblioteca digital armazena documentos digitais ou sua localização (Web ou física) e seus respectivos metadados, permitindo que um agente de um provedor de serviços faça colheita dos metadados de tais documentos, através do protocolo OAI-PMH. Nosso sistema lida com metadados de documentos que estão descritos em XML no padrão DC.

Assim, os dados utilizados como fonte para a tarefa de recomendação consistem de: (i) informações do usuário, obtidas a partir do currículo Lattes em XML; e (ii) informações sobre os documentos digitais, obtidas através de metadados no formato DC codificados em XML.

O padrão XML para o *Curriculum Vitae* (CV) da Plataforma Lattes é mantido pela Comunidade CONSCIENTIAS (Comunidade para Ontologias em Ciência, Tecnologia e Informações de Aperfeiçoamento de Nível Superior), uma extensão da comunidade LPML (Linguagem de Marcação da Plataforma Lattes). A gramática construída para tal padrão na linguagem de esquemas *XML Schema* do Consórcio W3C pode ser obtida em (LPML-CNPq, 2005), bem como sua documentação. O XML utilizado para descrever os documentos digitais, obtido como resultado do processo de colheita, segue o *XML Schema* apresentado em (DC-OAI, 2005).

A seguir serão apresentados: o perfil de usuário utilizado neste trabalho, a arquitetura do Sistema de Recomendação proposto e o modelo de recomendação adotado. Os detalhes de implementação do sistema também serão discutidos, sendo apresentados: a base de dados local e cada um dos módulos da arquitetura.

4.1 Perfil do usuário - Currículo Lattes

Considerando que um sistema de recomendação trata da personalização de informação, é essencial que este lide com o perfil do usuário. Em nosso trabalho, este perfil é obtido através do *Curriculum Vitae* do usuário, no caso o currículo Lattes é utilizado. O currículo Lattes é uma iniciativa do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). A plataforma Lattes oferece uma base de dados padrão dos currículos de pesquisadores e acadêmicos no Brasil. A plataforma é utilizada para: (i) avaliar a competência dos usuários e acadêmicos para garantias de concessões; (ii) para selecionar membros de comitês; e (iii) para auxiliar no processo de avaliação de projetos de pesquisa e cursos de pós-graduação.

A Tabela 4.1 mostra um subconjunto dos elementos de metadados do currículo Lattes. A mesma apresenta categorias de metadados, utilizadas em nosso trabalho para suportar o processo de recomendação, e suas descrições associadas. Para melhor compreensão, as categorias e seus metadados foram nomeados da forma apresentada na tabela, sendo que o prefixo "cv:" é usado, neste trabalho, para referenciar esses elementos de metadados do currículo Lattes.

Tabela 4.1: Subconjunto dos elementos de metadados do currículo Lattes

Categoria de Metadados	*Descrição*
Personal information	Esta categoria contém informações gerais sobre o usuário. Alguns metadados são: - *cv:name* - *cv:personal-address* - *cv:professional-address*
Academic graduation	Esta categoria contém informações do usuário sobre sua formação acadêmica. Alguns metadados são: - *cv:graduation-level* (*Undergraduate*, *Master graduate*, *PhD. graduate*) - *cv:graduation-year* - *cv:monograph-title* - *cv:monograph-keywords* - *cv:monograph-area* - *cv:monograph-advisor*
Language proficiency	Esta categoria contém informações sobre os idiomas que o usuário possui alguma proficiência. Alguns metadados são: - *cv:language* - *cv:language-skill* (*reading*, *writing*, *speaking*, *comprehension*) - *cv:language-skill-level* (*good*, *reasonable*, *little*)

Categoria de Metadados	*Descrição*
Bibliographic production	Esta categoria provê informações do usuário sobre suas produções bibliográficas em eventos, jornais, capítulos de livros, etc. Alguns metadados são: - *cv:article-title* - *cv:article-keywords* - *cv:article-language* - *cv:article-year*

4.2 Arquitetura do Sistema de Recomendação

Na Figura 4.1, é apresentada a arquitetura do provedor de serviços proposto neste trabalho, que oferece o serviço de recomendação de artigos científicos.

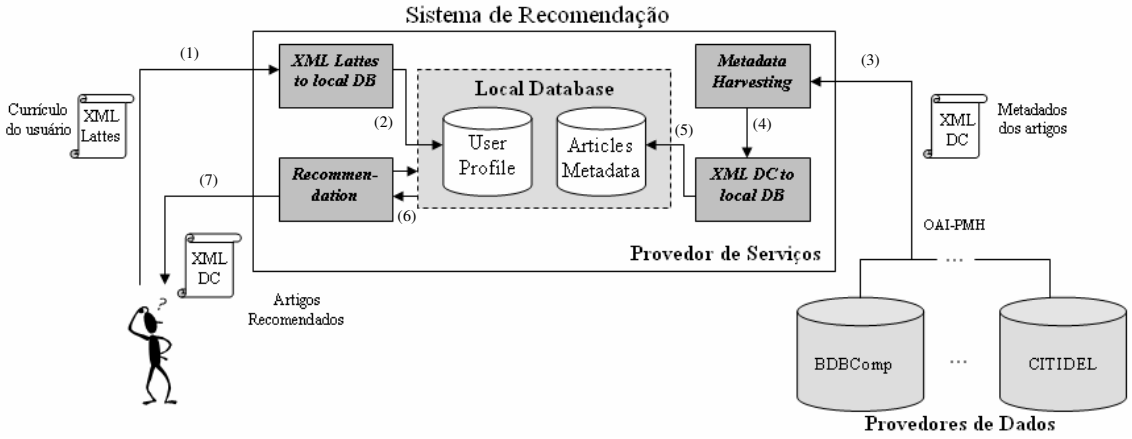

Figura 4.1: Arquitetura do Sistema

Para iniciar o processo de recomendação, o usuário deve fornecer ao sistema seu currículo Lattes em XML. Sempre que um usuário fizer o registro no sistema e carregar seu currículo Lattes (1), o módulo ***XML Lattes to local DB*** será ativado e a informação sobre os interesses do usuário será armazenada na base de dados local nomeada *User Profile* (2). O módulo ***Metadata Harvesting*** é ativado para enviar requisições a um provedor de dados, com o objetivo de fazer colheita (*harvesting*) de metadados de documentos digitais específicos de uma Biblioteca Digital. Ele recebe um documento XML como resposta (3) e o módulo ***XML DC to local DB*** é ativado (4). Este módulo extrai os metadados relevantes, para gerar a recomendação, de documentos XML no formato DC e armazena-os em uma base de dados local nomeada *Articles Metadata* (5). Uma vez que o perfil do usuário e os metadados dos artigos a serem recomendados estão disponíveis na base de dados local (*Local Database*), o módulo ***Recommendation*** pode ser ativado (6). O foco é recomendar artigos de uma Biblioteca Digital que melhor "casem" com o perfil do usuário descrito através do seu currículo Lattes (7).

A seguir será apresentado o modelo de recomendação implementado pelo módulo ***Recommendation***.

4.3 O modelo de recomendação

A recomendação é baseada no modelo vetorial. Neste caso, do perfil do usuário (currículo Lattes) serão extraídos os termos que irão compor o vetor de consulta para representar a necessidade de informação do usuário. O vetor de consulta é construído com termos obtidos de *cv:monograph-title* e *cv:monograph-keywords* da categoria *Academic graduation* (Tabela 4.1) e *cv:article-title* e *cv:article-keywords* da categoria *Bibliographic production* (Tabela 4.1) do currículo Lattes.

Para composição do vetor de consulta, das palavras contidas em *cv:monograph-title* e *cv:article-title* são eliminadas as *stopwords* (CLEF, 2005) (uma lista de termos comuns ou gerais, que não são usados no processo de recuperação de informação por possuírem conteúdo semântico limitado e serem muito freqüentes em todos os documentos, tais como: preposições, artigos e conjunções), sendo que cada palavra restante é considerada um termo simples. Por outro lado, em *cv:monograph-keywords* e *cv:article-keywords*, os termos são considerados integralmente, como expressões (termos simples ou compostos).

Os pesos do vetor de consulta são construídos de acordo com a Equação 4.1. Esta equação é obtida pelo produto de três pesos auxiliares: (i) $w_{KeywordOrTitle}$, que leva em consideração o tipo do termo (se obtido de "palavra-chave" ou "título"); (ii) $w_{Language}$, que leva em consideração o idioma do termo considerado; e (iii) w_{Year}, que leva em consideração o ano de conclusão/publicação (da formação acadêmica ou produção bibliográfica da qual o termo foi originado).

Termos obtidos de "palavras-chave" (*cv:monograph-keywords* e *cv:article-keywords*) são considerados mais importantes que os obtidos de "título" (*cv:monograph-title* e *cv:article-title*) e recebem um peso associado maior, isso se deve ao fato de "palavras-chave", geralmente, serem bastante representativas para indexar um trabalho, enquanto o "título" poderá ser menos relevante, já que este pode conter siglas e palavras não representativas. Assim, termos obtidos de publicações em idiomas, nos quais o usuário possui uma maior proficiência de leitura, são mais valorizados, pois o peso é atribuído de acordo com a mesma (*cv:language-skill-level* de *cv:language-skill="reading"*), peso em ordem crescente para "*little*", "*reasonable*" e "*good*". Além disso, a termos obtidos de cursos de formação acadêmica e produções bibliográficas mais recentes (*cv:graduation-year* e *cv:article-year*) são atribuídos pesos maiores que àqueles menos recentes; a idéia é de que informações mais recentes do currículo são mais relevantes, para determinar assuntos sobre os quais o usuário gostaria de receber algum tipo de recomendação no momento.

$$w_t = w_{KeywordOrTitle} . w_{Language} . w_{Year} \tag{4.1}$$

Os pesos auxiliares $w_{KeywordOrTitle}$, $w_{Language}$ e w_{Year} são calculados de acordo com a Equação 4.2.

$$w_i = 1 - (i-1).\left(\frac{1 - w_{\min}}{n-1}\right) \tag{4.2}$$

Nesta equação, os valores dos parâmetros, utilizados para cálculo de w_i, irão variar de acordo com o peso auxiliar que está sendo computado. Os parâmetros necessários neste cálculo são os seguintes: (i) n, que indica o número de possibilidades de pesos distintos que pode ser obtido para cálculo do peso auxiliar em questão; (ii) i, que indica

a possibilidade que está sendo calculada, sendo que seu valor pode variar de 1 a *n*; e (iii) w_{min}, que representa o valor mínimo de peso que pode ser obtido. Dessa forma, o valor de w_i obtido irá variar de 1 a w_{min}, para *i* variando de 1 a *n*.

Ilustrando, os valores dos parâmetros utilizados na avaliação experimental (que será apresentada no capítulo 5) são: (i) para $w_{KeywordOrTitle}$, w_{min} foi 0.95, *n* é 2, e *i* sendo 1 para termos obtidos de *keywords* e 2 para termos obtidos de *title*; (ii) para $w_{Language}$, w_{min} foi 0.60, *n* é 3, e *i* sendo 1 se o *cv:language-skill-level* é "*good*", 2 para "*reasonable*" e 3 para "*little*"; e (iii) para w_{Year}, w_{min} foi 0.55 e *i* varia de 1 até *n*, onde *n* é o intervalo de anos considerado, sendo 1 o maior e *n* o menor. Na avaliação experimental é considerado o intervalo entre 2006 e 2003, entretanto, se o intervalo for omitido, será considerado o intervalo entre o ano atual e o menor ano cadastrado no currículo do usuário (menor ano encontrado entre os anos de *cv:graduation-year* e *cv:article-year*).

Se w_{min} não for informado, o valor *default* será utilizado (apresentado na Equação 4.3). Nesta situação, a Equação 4.2 é reduzida à Equação 4.4.

$$w_{\min\, default} = \frac{1}{n} \tag{4.3}$$

$$w_i = \frac{n - i + 1}{n} \tag{4.4}$$

Uma vez que o vetor de consulta foi construído, é necessário calcular os pesos dos termos que o compõem, em cada um dos documentos que possivelmente serão recomendados, formando assim os vetores dos documentos. A abordagem adotada foi a *tf x idf* (*the product of the term frequency and the inverse document frequency*) (SALTON; BUCKLEY, 1988). Esta abordagem provê determinação automática dos pesos associados aos termos para recuperação de documentos. Na abordagem *tf x idf*, duas medidas são utilizadas em conjunto: *tf* e *idf*. *Term frequency* (*tf*) corresponde ao número de ocorrências de um termo no documento. *Inverse document frequency* (*idf*) (inverso da freqüência do termo na coleção) é um fator de varia inversamente ao número de documentos *n*, nos quais um termo é encontrado, de uma coleção de *N* documentos (tipicamente calculado como log (N/n)).

Os melhores termos, para identificação do conteúdo dos documentos, são aqueles capazes de distinguir documentos individuais do restante da coleção (SALTON et al., 1988). Assim, os melhores termos de indexação correspondem a termos com alta freqüência de aparecimento em um documento (alto valor de *tf*) e baixa freqüência de aparecimento na coleção (alto valor de *idf*). Para computar a métrica *tf x idf*, o sistema proposto utiliza informações contidas em metadados no formato DC, que descrevem os artigos que possivelmente serão recomendados, sendo os elementos *dc:title* e *dc:description* utilizados para representar o conteúdo dos documentos para o cálculo das métricas apresentadas. Além disso, como o sistema pode trabalhar com diferentes idiomas, o número total de documentos da coleção irá variar de acordo com o idioma do termo considerado. Determinados os vetores de consulta e de documentos, o sistema é capaz de computar os valores de similaridade entre os documentos e a consulta de acordo com a fórmula do modelo vetorial (Equação 3.2).

4.4 Implementação do sistema

Nesta seção será descrita, em maiores detalhes, a implementação do sistema de recomendação desenvolvido neste trabalho, para gerar recomendações de artigos científicos da área da Ciência da Computação. Artigos estes indexados por qualquer Biblioteca Digital compatível com o padrão OAI, que disponibilize seus metadados no formato DC via protocolo OAI-PMH. Sendo que o perfil do usuário, utilizado para gerar a recomendação personalizada, é o currículo Lattes em XML. Tal sistema trata-se de um provedor de serviços capaz de realizar colheita de metadados em Bibliotecas Digitais e gerar um serviço específico com base nos metadados obtidos, no caso, um serviço de recomendação personalizada.

Para que fosse possível a implementação deste sistema, foi necessária a utilização de uma linguagem de programação para Web, que permitisse a criação de páginas dinâmicas e oferecesse suporte para acessos a bancos de dados. Para a implementação do provedor de serviços, foi utilizada a linguagem PHP (PHP Group, 2006) que oferece funções para interação com diversos bancos de dados, dentre eles o banco de dados MySQL (MySQL, 2006), que foi utilizado para implementação da base de dados local, além de oferecer funções para trabalhar com documentos XML (NIEDERAUER, 2002). O banco de dados MySQL e a linguagem de programação PHP também foram escolhidos por serem gratuitos, de código aberto e portáveis (podendo ser utilizados em sistemas operacionais com Windows e Linux).

4.4.1 *Local Database*

A base de dados local, como apresentada na arquitetura (Figura 4.1), é constituída de informações relevantes para o processo de recomendação, sobre duas entidades principais: usuários (*User profile*) e artigos (*Articles Metadata*). As modelagens ER, contendo as tabelas da base de dados relacionadas com a descrição de cada uma dessas entidades, são apresentadas, respectivamente, na Figura 4.2 e na Figura 4.3. Tal separação foi realizada apenas para facilitar a compreensão e a visualização das informações. A base foi implementada utilizando o banco de dados relacional MySQL.

A base de dados *User profile* é carregada com as informações do perfil do usuário extraídas do currículo Lattes dos mesmos. Já na base de dados *Metadata Articles*, algumas tabelas são carregadas com os metadados DC, relevantes para o processo de recomendação, de arquivos XML obtidos através de colheita de metadados de artigos de determinadas edições de conferências da área da Ciência da Computação, indexados em Bibliotecas Digitais. A tabela *stopword* contém a lista de *stopwords* utilizada pelo sistema por idioma. As tabelas *term* e *terms_articles* têm suas informações geradas automaticamente pelo sistema, para facilitar o processo de recomendação realizado.

Os arquivos XML, utilizados como fonte de dados pelo sistema, são catalogados, havendo referência para a localização do arquivo físico no servidor em que o sistema de recomendação está instalado. Dessa forma, os metadados não cadastrados na base, podem ser recuperados a qualquer momento para uma descrição mais completa de usuários e artigos a serem recomendados. Cabe observar que a tabela *language* está representada duas vezes, nos dois modelos ER apresentados, apenas para facilitar a compreensão.

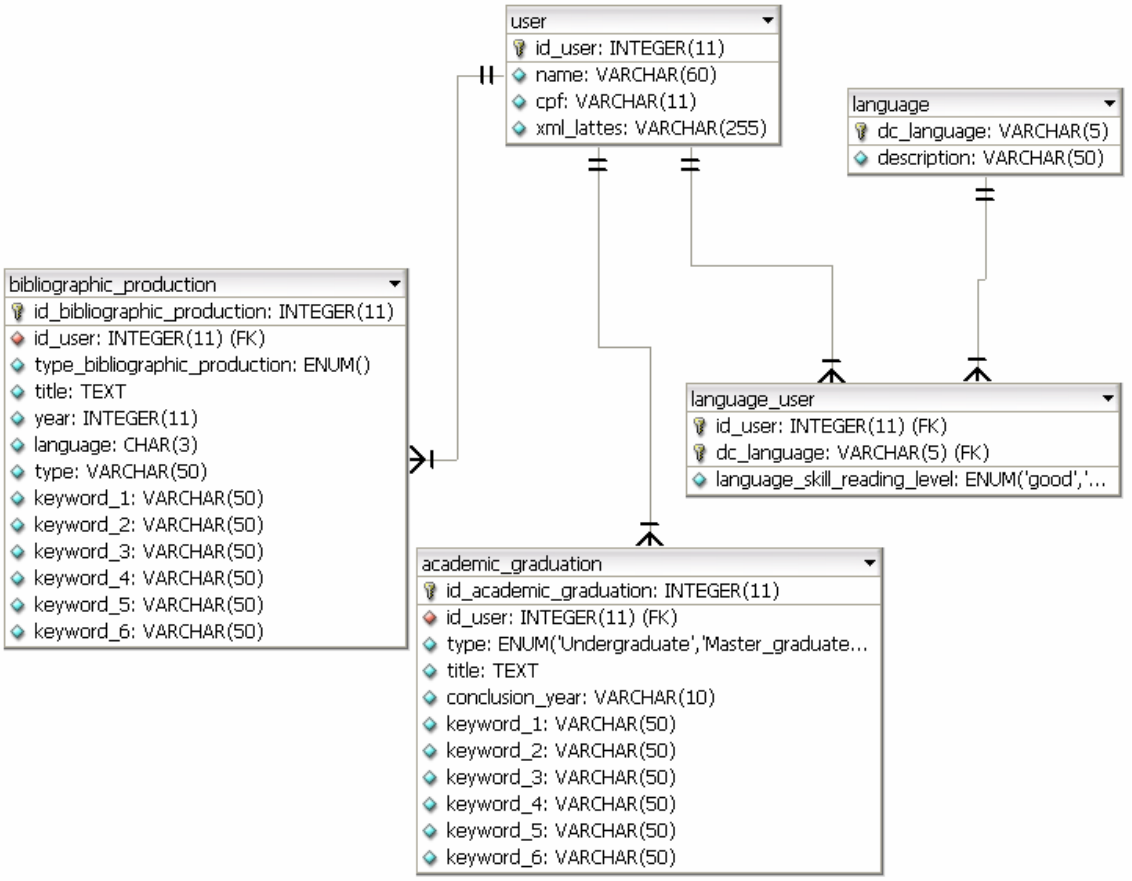

Figura 4.2: Modelo ER da base de dados local *User Profile*

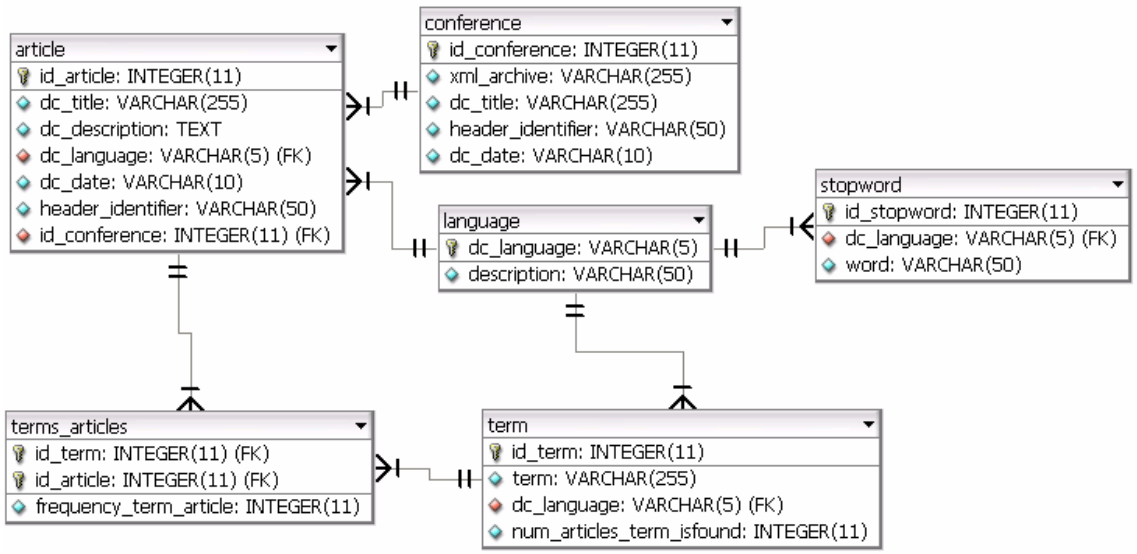

Figura 4.3: Modelo ER da base de dados local *Articles Metadata*

4.4.2 *XML Lattes to local DB*

O módulo *XML Lattes to local DB* possui a função de extrair as informações do currículo Lattes do usuário em XML, que são consideradas relevantes para a tarefa de recomendação personalizada de artigos científicos, e armazená-las na base de dados

local (*User profile*). Tais informações correspondem ao subconjunto do currículo Lattes apresentado na Tabela 4.1. Na implementação deste módulo, são utilizadas as funções de PHP, que operam sobre arquivos XML, havendo a utilização de um parser SAX (*Simple API for XML*) e funções para interação com o banco de dados MySQL.

Na Figura 4.4, é apresentado um trecho exemplo de um currículo Lattes no formato XML. Tem-se como raiz a tag CURRICULO-VITAE e alguns atributos desta tag que irão identificar se o arquivo correspondente é de fato um arquivo XML do currículo Lattes. Neste exemplo, podem ser visualizadas algumas tags e atributos contendo informações relevantes para o processo de recomendação.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<CURRICULO-VITAE SISTEMA-ORIGEM-XML="LATTES_OFFLINE" DATA-
ATUALIZACAO="24102006" HORA-ATUALIZACAO="182818"
xmlns:lattes="http://www.cnpq.br/2001/XSL/Lattes">
<DADOS-GERAIS NOME-COMPLETO="Giseli Rabello Lopes" NOME-EM-CITACOES-
BIBLIOGRAFICAS="LOPES, Giseli Rabello" NACIONALIDADE="B"
CPF="00000000000" PAIS-DE-NASCIMENTO="Brasil"...>
...
<FORMACAO-ACADEMICA-TITULACAO>
...
<MESTRADO SEQUENCIA-FORMACAO="9" NIVEL="3" CODIGO-
INSTITUICAO="019200000005" NOME-INSTITUICAO="Universidade Federal do
Rio Grande do Sul" CODIGO-CURSO="42000041" NOME-CURSO="Ciência da
Computação" CODIGO-AREA-CURSO="10300007" STATUS-DO-
CURSO="EM_ANDAMENTO" ANO-DE-INICIO="2005" ANO-DE-CONCLUSAO="" FLAG-
BOLSA="SIM" CODIGO-AGENCIA-FINANCIADORA="002200000000" NOME-
AGENCIA="Conselho Nacional de Desenvolvimento Científico e
Tecnológico" ANO-DE-OBTENCAO-DO-TITULO="" TITULO-DA-DISSERTACAO-
TESE="Sistema de Recomendação para Bibliotecas Digitais sob a
Perspectiva da Web Semântica" NOME-COMPLETO-DO-ORIENTADOR="José
Palazzo Moreira de Oliveira">
<PALAVRAS-CHAVE PALAVRA-CHAVE-1="Sistemas de Recomendação" PALAVRA-
CHAVE-2="Bibliotecas Digitais" PALAVRA-CHAVE-3="Personalização da
Informação" PALAVRA-CHAVE-4="Modelo Vetorial" PALAVRA-CHAVE-
5="Provedor de Serviços" PALAVRA-CHAVE-6="Web Semântica" />
...
</MESTRADO>
</FORMACAO-ACADEMICA-TITULACAO>
...
<IDIOMAS>
<IDIOMA IDIOMA="EN" DESCRICAO-DO-IDIOMA="Inglês" PROFICIENCIA-DE-
LEITURA="BEM".../>
...
</IDIOMAS>
...
</DADOS-GERAIS>
...
<PRODUCAO-BIBLIOGRAFICA>
<TRABALHOS-EM-EVENTOS>
...
<TRABALHO-EM-EVENTOS SEQUENCIA-PRODUCAO="38"><DADOS-BASICOS-DO-
TRABALHO NATUREZA="COMPLETO" TITULO-DO-TRABALHO="Sistema de
Recomendação para Bibliotecas Digitais sob a Perspectiva da Web
Semântica" ANO-DO-TRABALHO="2006" PAIS-DO-EVENTO="Brasil"
IDIOMA="Português" MEIO-DE-DIVULGACAO="IMPRESSO" HOME-PAGE-DO-
TRABALHO="" FLAG-RELEVANCIA="NAO" /><DETALHAMENTO-DO-TRABALHO
CLASSIFICACAO-DO-EVENTO="NACIONAL" NOME-DO-EVENTO="II Workshop de
Bibliotecas Digitais, WDL; SBBD/SBES" CIDADE-DO-
```

```
EVENTO="Florianópolis" ANO-DE-REALIZACAO="2006" TITULO-DOS-ANAIS-OU-
PROCEEDINGS="" VOLUME="" FASCICULO="" SERIE="" PAGINA-INICIAL="21"
PAGINA-FINAL="30" ISBN="8576690896" NOME-DA-EDITORA="Sociedade
Brasileira de Computação" CIDADE-DA-EDITORA="" /><AUTORES NOME-
COMPLETO-DO-AUTOR="Maria Aparecida Martins Souto" NOME-PARA-
CITACAO="SOUTO, Maria Aparecida Martins" ORDEM-DE-AUTORIA="2"
/><AUTORES NOME-COMPLETO-DO-AUTOR="José Palazzo Moreira de Oliveira"
NOME-PARA-CITACAO="OLIVEIRA, José Palazzo Moreira de" ORDEM-DE-
AUTORIA="3" /><AUTORES NOME-COMPLETO-DO-AUTOR="Giseli Rabello Lopes"
NOME-PARA-CITACAO="LOPES, Giseli Rabello" ORDEM-DE-AUTORIA="1" />
<PALAVRAS-CHAVE PALAVRA-CHAVE-1="Sistemas de Recomendação" PALAVRA-
CHAVE-2="Bibliotecas Digitais" PALAVRA-CHAVE-3="Tecnologias de
Personalização" PALAVRA-CHAVE-4=" " PALAVRA-CHAVE-5=" " PALAVRA-
CHAVE-6=" " />
<AREAS-DO-CONHECIMENTO><AREA-DO-CONHECIMENTO-1 NOME-GRANDE-AREA-DO-
CONHECIMENTO="CIENCIAS_EXATAS_E_DA_TERRA" NOME-DA-AREA-DO-
CONHECIMENTO="Ciência da Computação" NOME-DA-SUB-AREA-DO-
CONHECIMENTO="Sistemas de Informação" NOME-DA-ESPECIALIDADE=""
/></AREAS-DO-CONHECIMENTO>
...
</TRABALHO-EM-EVENTOS>
...
</TRABALHOS-EM-EVENTOS>
...
</PRODUCAO-BIBLIOGRAFICA>
...
</CURRICULO-VITAE>
```

Figura 4.4: Trecho de um currículo Lattes em XML

A Tabela 4.2 indica as categorias do subconjunto do currículo Lattes (apresentado na Tabela 4.1) e sua equivalência às tags do arquivo XML do currículo Lattes e às tabelas da base de dados local *User profile* (apresentada na Figura 4.2).

Tabela 4.2: Equivalência entre subconjunto de elementos de metadados do currículo Lattes, Tags no arquivo XML do currículo Lattes e tabelas da base de dados local *User Profile*

Categoria de Metadados do subconjunto do currículo Lattes	*Tag no arquivo XML do currículo Lattes*	*Tabela da base de dados local "User profile"*
Personal information	DADOS-GERAIS	user
Academic graduation	FORMACAO-ACADEMICA-TITULACAO	academic_graduation
Language proficiency	IDIOMAS	language_user
Bibliographic production	PRODUCAO-BIBLIOGRAFICA	bibliographic_production

O trecho do arquivo XML apresentado na Figura 4.4 está representando um subconjunto do currículo Lattes de *Giseli Rabello Lopes*, uma pessoa de nacionalidade brasileira, que possui um curso de formação acadêmica em andamento, no caso é um curso de Mestrado, cujo título da dissertação é *Sistema de Recomendação para Bibliotecas Digitais sob a Perspectiva da Web Semântica*. À descrição da dissertação estão associadas seis palavras-chave (*Sistemas de Recomendação*, *Bibliotecas Digitais*, *Personalização da Informação*, *Modelo Vetorial*, *Provedor de Serviços*, *Web Semântica*).

Também é representada uma produção bibliográfica de autoria da mesma pessoa, cujo título é *Sistema de Recomendação para Bibliotecas Digitais sob a Perspectiva da Web Semântica*, correspondente a um trabalho publicado em anais de um evento e que possui três palavras-chave (*Sistemas de Recomendação*, *Bibliotecas Digitais*, *Tecnologias de Personalização*) associadas à descrição desta publicação.

A seguir, são apresentados exemplos de mapeamentos de trechos do currículo Lattes em XML, da Figura 4.4, para a base de dados local *User Profile*. Nos exemplos, à esquerda é apresentado um trecho do arquivo XML com as tags e atributos de onde são obtidas as informações, que serão armazenadas nas tabelas da base de dados local e, à direita, são apresentadas as instruções SQL geradas para inserção dos dados apresentados no exemplo nas tabelas correspondentes. Nos exemplos, os identificadores das tabelas não informados são do tipo inteiro e com valores obtidos de auto-incremento (1 a *n*).

Na Figura 4.5, é apresentado um exemplo do mapeamento dos dados contidos pela tag *DADOS-GERAIS* para a tabela *user*. O usuário deste exemplo foi o primeiro inserido na base, tendo o valor de *id_user* igual a 1. Pode ser observado que o sistema infere que o usuário possui uma proficiência de leitura "bom" para sua língua materna, já que no currículo Lattes, a informação de proficiência de leitura para o idioma materno não é realizada de forma explícita pelo usuário.

<DADOS-GERAIS NOME-COMPLETO="Giseli Rabello Lopes" NOME-EM-CITACOES-BIBLIOGRAFICAS="LOPES, Giseli Rabello" NACIONALIDADE="B" CPF="00000000000" PAIS-DE-NASCIMENTO="Brasil"...>	insert into user (name, cpf, xml_lattes) values ('Giseli Rabello Lopes ', '00000000000', 'lattes_giseli.xml') Sendo user(id_user='1') e language(dc_language='por'; description='Português'): insert into language_user (id_user, dc_language, language_skill_reading_level) values ('1', 'por', 'good')

Figura 4.5: Mapeamento dos dados da tag *DADOS-GERAIS* para tabela *user*

Na Figura 4.6, é apresentado um exemplo do mapeamento dos dados contidos pela tag *FORMACAO-ACADEMICA-TITULACAO* para a tabela *academic_graduation*. Os cursos de formação acadêmica do usuário podem ser: graduação, mestrado ou doutorado (aqueles que possuem informações sobre o trabalho desenvolvido como: título e palavras-chave relacionadas à monografia, dissertação ou tese). Dessa forma, a estrutura do XML irá variar um pouco, de acordo com o curso representado. No caso deste exemplo, é apresentado um trecho em XML o qual descreve um curso de mestrado. Cabe ressaltar que, para cursos de formação acadêmica em andamento, o ano de conclusão (*conclusion_year*) é indicado como 0. O foco das recomendações, no caso do sistema proposto, é de artigos da área da Ciência da Computação. Sendo assim, somente informações do Lattes, referentes a esta área de pesquisa, serão armazenadas na base de dados local (isso pode ser identificado através do atributo CODIGO-AREA-CURSO).

```
<FORMACAO-ACADEMICA-TITULACAO>
...
<MESTRADO CODIGO-
CURSO="42000041" NOME-
CURSO="Ciência da Computação"
CODIGO-AREA-CURSO="10300007"
STATUS-DO-CURSO="EM_ANDAMENTO"
ANO-DE-CONCLUSAO="" TITULO-DA-
DISSERTACAO-TESE="Sistema de
Recomendação para Bibliotecas
Digitais sob a Perspectiva da
Web Semântica"...>
<PALAVRAS-CHAVE PALAVRA-CHAVE-
1="Sistemas de Recomendação"
PALAVRA-CHAVE-2="Bibliotecas
Digitais" PALAVRA-CHAVE-
3="Personalização da Informação"
PALAVRA-CHAVE-4="Modelo
Vetorial" PALAVRA-CHAVE-
5="Provedor de Serviços"
PALAVRA-CHAVE-6="Web Semântica"
/>
...
</MESTRADO>
</FORMACAO-ACADEMICA-TITULACAO>
```

```
insert into academic_graduation
(id_user, type, title,
conclusion_year, keyword_1,
keyword_2, keyword_3, keyword_4,
keyword_5, keyword_6) values ('1',
'Master graduate', 'Sistema de
Recomendação para Bibliotecas
Digitais sob a Perspectiva da Web
Semântica', '0', 'Sistemas de
Recomendação', 'Bibliotecas
Digitais', 'Personalização da
Informação', 'Modelo Vetorial',
'Provedor de Serviços', 'Web
Semântica')
```

Figura 4.6: Mapeamento dos dados da tag *FORMACAO-ACADEMICA-TITULACAO* para tabela *academic_graduation*

Na Figura 4.7, é apresentado um exemplo do mapeamento dos dados contidos pela tag *IDIOMAS* para a tabela *language_user*. Para o processo de recomendação, apenas a proficiência de leitura é levada em consideração, portanto, somente esta informação é armazenada na base de dados.

```
<IDIOMAS>
<IDIOMA IDIOMA="EN" DESCRICAO-
DO-IDIOMA="Inglês" PROFICIENCIA-
DE-LEITURA="BEM".../>
...
</IDIOMAS>
```

Sendo language(dc_language='eng'; description='Inglês'):

```
insert into language_user
(id_user, dc_language,
language_skill_reading_level)
values ('1', 'eng', 'good')
```

Figura 4.7: Mapeamento dos dados da tag *IDIOMAS* para tabela *language_user*

Por fim, na Figura 4.8, é apresentado um exemplo do mapeamento dos dados contidos pela tag *PRODUCAO-BIBLIOGRAFICA*, para a tabela *bibliographic_production*. Os tipos de produção bibliográfica do usuário podem ser: "trabalho em evento", "artigo publicado", "texto em jornal ou revista", "livro publicado ou organizado" e "capítulo de livro publicado". Dessa forma, a estrutura do XML irá variar um pouco, de acordo com o tipo de publicação. No exemplo dado, é apresentado um trecho em XML, que descreve um trabalho em evento. Na base de dados local, as informações relevantes, para o processo de recomendação obtidas de tais produções bibliográficas, são todas armazenadas na tabela *bibliographic_production*. No caso da produção bibliográfica, para identificar, se a área de pesquisa do trabalho é "Ciência da Computação", existem áreas do conhecimento associadas a cada publicação. Infelizmente, esta informação não é de preenchimento obrigatório no currículo Lattes. Assim, para compor a recomendação são levados em consideração trabalhos que

possuam área do conhecimento (NOME-DA-AREA-DO-CONHECIMENTO) informada como "Ciência da Computação" ou, caso esta informação não esteja disponibilizada, o trabalho é igualmente considerado.

```
<PRODUCAO-BIBLIOGRAFICA>
<TRABALHOS-EM-EVENTOS>
...
<TRABALHO-EM-EVENTOS...> <DADOS-BASICOS-DO-TRABALHO NATUREZA="COMPLETO" TITULO-DO-TRABALHO="Sistema de Recomendação para Bibliotecas Digitais sob a Perspectiva da Web Semântica" ANO-DO-TRABALHO="2006" IDIOMA="Português".../>
...
<PALAVRAS-CHAVE PALAVRA-CHAVE-1="Sistemas de Recomendação" PALAVRA-CHAVE-2="Bibliotecas Digitais" PALAVRA-CHAVE-3="Tecnologias de Personalização" PALAVRA-CHAVE-4=" " PALAVRA-CHAVE-5=" " PALAVRA-CHAVE-6=" " />
<AREAS-DO-CONHECIMENTO><AREA-DO-CONHECIMENTO-1 NOME-GRANDE-AREA-DO-CONHECIMENTO="CIENCIAS_EXATAS_E_DA_TERRA" NOME-DA-AREA-DO-CONHECIMENTO="Ciência da Computação" NOME-DA-SUB-AREA-DO-CONHECIMENTO="Sistemas de Informação" NOME-DA-ESPECIALIDADE="" /></AREAS-DO-CONHECIMENTO>
...
</TRABALHO-EM-EVENTOS>
...
</TRABALHOS-EM-EVENTOS>
...
</PRODUCAO-BIBLIOGRAFICA>
```

```
insert into
bibliographic_production (id_user,
type_bibliographic_production,
title, year, language, type,
keyword_1,  keyword_2, keyword_3)
values ('1',  'publication in
proceedings', 'Sistema de
Recomendação para Bibliotecas
Digitais sob a Perspectiva da Web
Semântica', '2006', 'por',
'complete', 'Sistemas de
Recomendação', 'Bibliotecas
Digitais', 'Tecnologias de
Personalização')
```

Figura 4.8: Mapeamento dos dados da tag *PRODUCAO_BIBLIOGRAFICA* para tabela *bibliographic_production*

4.4.3 *Metadata Harvesting*

O módulo *Metadata Harvesting* é o *harvester* do provedor de serviços implementado neste trabalho. Esta é uma aplicação cliente que envia requisições OAI-PMH (ver seção 2.1.1 para detalhes sobre as requisições possíveis), para realizar a colheita de metadados de provedores de dados de Bibliotecas Digitais. Na Figura 4.9, é apresentado um trecho de código de um possível *harvester* de metadados para a BDBComp. A variável $arq conterá a requisição completa desejada. No caso, temos a URL do provedor de dados da BDBComp "http://www.lbd.dcc.ufmg.br/cgi-bin/bdbcomp/oai2/oai.pl", o verbo OAI-PMH "*ListRecords*" (para listar os metadados dos registros), cujo *metadataPrefix* é "*oai_dc*" (que corresponde ao formato de metadados Dublin Core) e o conjunto desejado (*set*). Neste caso, a requisição será

repetida para todos os conjuntos da BDBComp, sendo cada conjunto, equivalente a uma edição de determinada conferência. No código exemplo, os arquivos XML, recebidos como resposta, serão armazenados no diretório "BDBComp_xml_articles", cujos nomes correspondem aos números dos conjuntos correspondentes (variação de $x).

```
<?
for ($x=1;$x<=$nro_set_maximo;$x++)
{
    $arq="http://www.lbd.dcc.ufmg.br/cgi-
bin/bdbcomp/oai2/oai.pl?verb=ListRecords&metadataPrefix=oai_dc&set=".
$x;
    $dest="BDBComp_xml_articles/".$x.".xml";
    $lines = file($arq);
    $fp = fopen($dest, 'w+');
    foreach($lines as $line)
        fwrite($fp, $line);
    fclose($fp);
}
echo "Colheita de metadados da BDBComp finalizada...";
?>
```

Figura 4.9: Trecho de código de um *harvester* para a BDBComp

A seguir, na Figura 4.10, é apresentado um trecho do XML recebido como resposta à requisição "http://www.lbd.dcc.ufmg.br/cgi-bin/bdbcomp/oai2/oai.pl?verb= ListRecords&metadataPrefix=oai_dc&set=109". Esta requisição solicita os metadados em DC, que descrevem os artigos publicados no SBBD (Simpósio Brasileiro de Banco de Dados) de 2005, representado na BDBComp pelo *set=109*.

```
   <?xml version="1.0" encoding="UTF-8"?>

   <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">

   <responseDate>2006-05-31T16:04:33Z</responseDate>
   <request verb="ListRecords" metadataPrefix="oai_dc"
set="109">http://www.lbd.dcc.ufmg.br/cgi-
bin/bdbcomp/oai2/oai.pl</request>

   <ListRecords>
   <record>
   <header>
   <identifier>sbbd2005meta</identifier>
   <datestamp>2005-09-08</datestamp>
   <setSpec>109</setSpec>
   </header>
   <metadata>
   <oaidc:dc xmlns="http://purl.org/dc/elements/1.1/"
xmlns:oaidc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
```

```
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
   <title>XX Simpósio Brasileiro de Banco de Dados</title>
   <creator>Carlos A. Heuser</creator>
   <date>2005</date>
   <type>Collection</type>
   <identifier>sbbd2005</identifier>
   <language>por</language>
   <coverage>Uberlândia, MG, Brasil</coverage>
   <rights>Sociedade Brasileira de Computação</rights>
   </oaidc:dc>
   </metadata>
   </record>
   <record>
   <header>
   <identifier>sbbd2005meta001</identifier>
   <datestamp>2005-09-08</datestamp>
   <setSpec>109</setSpec>
   </header>
   <metadata>
   <oaidc:dc xmlns="http://purl.org/dc/elements/1.1/"
xmlns:oaidc="http://www.openarchives.org/OAI/2.0/oai_dc/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
   <title>Self Describing Components: Searching for Digital Artifacts
on the Web</title>
   <creator>André Santanchè</creator>
   <creator>Claudia Bauzer Medeiros</creator>
   <description>The Semantics Web has opened new horizons in
exploring Web functionality. One of the many challenges is to
proactively support the reuse of digital artifacts stored in
repositories all over the world. Our goal is to contribute towards
this issue, proposing a mechanism for describing and discovering
artifacts called Digital Content Components (DCCs). DCCs are self-
contained stored entities that may comprise any digital content, such
as pieces of software, multimedia or text. Their specification takes
advantage of Semantic Web standards and ontologies, both of which are
used in the discovery process. DCC construction and composition
procedures naturally lend themselves to patternmatching and
subsumption-based search. Thus, many existing methods for Web
searching can be extended to look for reusable artifacts. We validate
the proposal discussing its implementation for agro-environmental
planning.</description>
   <date>2005</date>
   <type>Text</type>
   <identifier>sbbd2005article001</identifier>
   <identifier>http://www.sbbd-sbes2005.ufu.br/arquivos/artigo-01-
BauzerSantache.pdf</identifier>
   <source>sbbd2005</source>
   <language>eng</language>
   <coverage>Uberlândia, MG, Brasil</coverage>
   <rights>Sociedade Brasileira de Computação</rights>
   </oaidc:dc>
   </metadata>
   </record>
   ...
   </ListRecords>

   </OAI-PMH>
```

Figura 4.10: Trecho de um arquivo XML respondendo à requisição *ListRecords*

4.4.4 *XML DC to local DB*

O módulo *XML DC to local DB* é responsável por interpretar os arquivos XML, recebidos pelo *harvester* (seção 4.4.3), e armazenar as informações que são utilizadas durante o processo de recomendação, presentes nestes, na base de dados local *Articles Metadata* (ver Figura 4.3). Estes metadados devem estar descritos no formato Dublin Core (ver Tabela 2.1 para detalhes sobre este formato). Este módulo, a exemplo do módulo *XML Lattes to local DB*, também foi implementado em PHP utilizando funções que operam sobre arquivos XML, sendo utilizado um parser SAX (*Simple API for XML*) e funções para interação com o banco de dados MySQL.

A seguir, são demonstrados exemplos de mapeamento dos dados do trecho do arquivo XML apresentado na Figura 4.10 para a base de dados local *Articles Metadata* (a forma de apresentação dos exemplos é a mesma adotada na seção 4.4.2). Nos arquivos XML, recebidos como resposta a requisições *ListRecords* por *sets* da BDBComp, o primeiro registro irá conter os metadados que descrevem a edição da conferência em que tais artigos foram publicados (dados estes que devem ser armazenados na tabela *conference* da base) e os registros seguintes correspondem aos metadados que descrevem os artigos propriamente ditos (dados estes que devem ser armazenados na tabela *article*) (ver Tabela 2.2 para detalhes sobre cada um dos elementos do formato DC na BDBComp). Dessa forma, o exemplo da Figura 4.11 mostra como é feito o mapeamento dos dados do primeiro registro retornado pelo *ListRecords* para a tabela *conference*. Já na Figura 4.12, temos o mapeamento dos dados do segundo registro retornado pelo *ListRecords* para a tabela *article*.

```
<record>
<header>
<identifier>sbbd2005meta</identi
fier>
...
<setSpec>109</setSpec>
</header>
<metadata>
<oaidc:dc ...>
<title>XX Simpósio Brasileiro de
Banco de Dados</title>
...
<date>2005</date>
...
<language>por</language>
...
</oaidc:dc>
</metadata>
</record>
```

```
insert into
conference(xml_archive, dc_title,
header_identifier, dc_date) values
('109.xml', 'XX Simpósio
Brasileiro de Banco de Dados',
'sbbd2005meta', '2005')
```

Sendo (id_conference=1)

Figura 4.11: Mapeamento dos dados para a tabela *conference*

Figura 4.12: Mapeamento dos dados para a tabela *article*

4.4.5 *Recommendation*

Com base nas informações armazenadas na base local, referentes ao currículo do usuário, e também naquelas referentes aos artigos, o módulo *Recommendation* realiza a tarefa de recomendação propriamente dita. O modelo de recomendação implementado neste sistema está descrito em detalhes na seção 4.3 deste trabalho. Este módulo é implementado em PHP e utiliza funções para permitir a interação com o banco de dados MySQL.

Na Figura 4.13, é apresentado um esquema de como são geradas as recomendações pelo módulo *Recommendation*. O primeiro passo consiste da determinação dos termos que irão compor o vetor de consulta que representa a necessidade de informação do usuário, tais informações são obtidas da base de dados local *User Profile* (1). Além disso, a cada termo é associado um peso indicando sua importância para a consulta, tal peso é calculado utilizando a equação apresentada na figura (Equação 4.1) (2). Após ter sido determinado o vetor de consulta, o peso dos termos que o compõem precisa ser determinado nos documentos a serem recomendados (formando os vetores dos documentos), isto é feito utilizando a abordagem *tf x idf* apresentada na seção 4.3; as informações sobre os documentos são obtidas da base local *Articles Metadata* (3). Tendo tais vetores determinados, é possível calcular a similaridade entre os vetores dos documentos e o vetor de consulta, através da fórmula do modelo vetorial (Equação 3.2) (4). Calculados os valores de similaridade entre a consulta e cada um dos documentos os resultados podem ser ordenados em ordem decrescente, formando o *ranking* dos documentos (quanto maior o valor encontrado, maior será a similaridade do documento com a consulta definida) (5). Por fim, as recomendações podem ser apresentadas ao usuário de acordo com o *ranking* obtido (6).

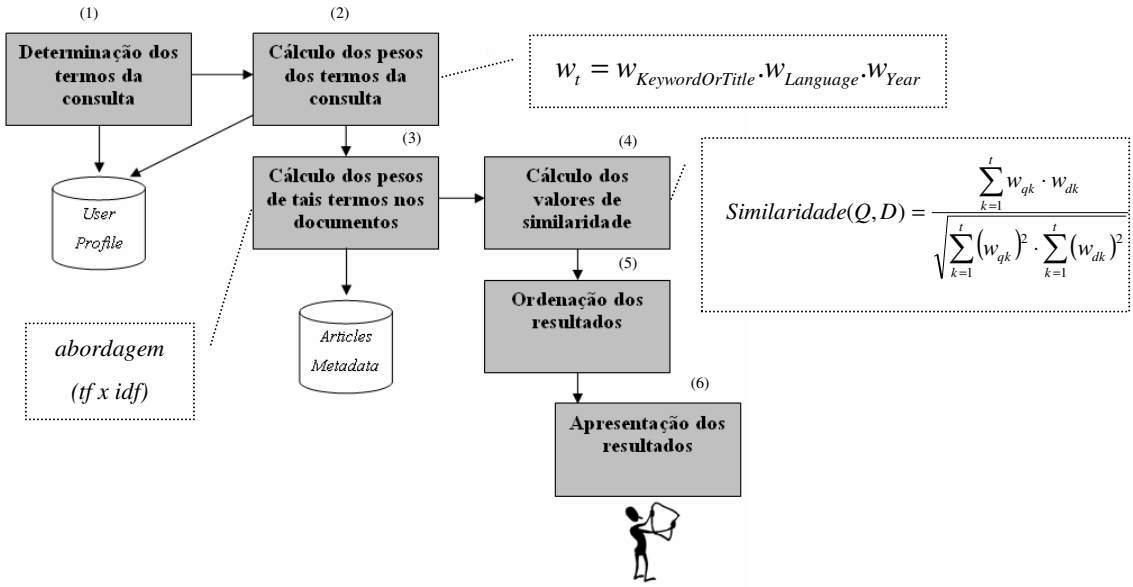

Figura 4.13: Esquema do módulo *Recommendation*

A seguir, será apresentado um exemplo da montagem de um vetor de consulta (passos (1) e (2) da Figura 4.13). Tal vetor representa os interesses de informação do usuário, que servirá de base para o processo de geração das recomendações. Na Figura 4.14, é apresentado um currículo Lattes em XML de um usuário de teste. Nesta figura, encontram-se em negrito algumas informações, relevantes para o processo de recomendação, que devem ser armazenadas na base de dados local *User Profile*.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<CURRICULO-VITAE SISTEMA-ORIGEM-XML="LATTES_OFFLINE" DATA-
ATUALIZACAO="24102006" HORA-ATUALIZACAO="182818"
xmlns:lattes="http://www.cnpq.br/2001/XSL/Lattes">

<DADOS-GERAIS NOME-COMPLETO="Usuário de Teste" NOME-EM-CITACOES-
BIBLIOGRAFICAS="TESTE, Usuário de" NACIONALIDADE="B"
CPF="99999999999" PAIS-DE-NASCIMENTO="Brasil">

<IDIOMAS>
<IDIOMA IDIOMA="EN" DESCRICAO-DO-IDIOMA="Inglês" PROFICIENCIA-DE-
LEITURA="BEM"/>
</IDIOMAS>

</DADOS-GERAIS>

<PRODUCAO-BIBLIOGRAFICA>

<TRABALHOS-EM-EVENTOS>

<TRABALHO-EM-EVENTOS SEQUENCIA-PRODUCAO="1">
<DADOS-BASICOS-DO-TRABALHO NATUREZA="COMPLETO" TITULO-DO-
TRABALHO="Bibliographic Recommendations for individual users" ANO-DO-
TRABALHO="2005" PAIS-DO-EVENTO="Brasil" IDIOMA="Inglês" MEIO-DE-
DIVULGACAO="IMPRESSO" HOME-PAGE-DO-TRABALHO="" FLAG-RELEVANCIA="NAO"
/>
<DETALHAMENTO-DO-TRABALHO CLASSIFICACAO-DO-EVENTO="NACIONAL" NOME-DO-
EVENTO="Evento Nacional" CIDADE-DO-EVENTO="Porto Alegre" ANO-DE-
REALIZACAO="2005" TITULO-DOS-ANAIS-OU-PROCEEDINGS="" VOLUME=""
FASCICULO="" SERIE="" PAGINA-INICIAL="01" PAGINA-FINAL="10"
ISBN="0000000000" NOME-DA-EDITORA="Sociedade Brasileira de
```

```
Computação" CIDADE-DA-EDITORA="" />
<AUTORES NOME-COMPLETO-DO-AUTOR="Usuário de Teste" NOME-PARA-
CITACAO="TESTE, Usuário de" ORDEM-DE-AUTORIA="1" />
<PALAVRAS-CHAVE PALAVRA-CHAVE-1="Recommender Systems" PALAVRA-CHAVE-
2="Collaborative Filtering" PALAVRA-CHAVE-3=" " PALAVRA-CHAVE-4=" "
PALAVRA-CHAVE-5=" " PALAVRA-CHAVE-6=" " />

</TRABALHO-EM-EVENTOS>

</TRABALHOS-EM-EVENTOS>

</PRODUCAO-BIBLIOGRAFICA>

</CURRICULO-VITAE>
```

Figura 4.14: Arquivo XML do currículo Lattes do *Usuário de Teste*

Com base nas informações sobre este usuário, que foram armazenadas na base de dados local, o sistema de recomendação poderá montar o vetor de consulta, que representará a necessidade de informação do usuário e que conterá os termos, que serão buscados nos documentos a serem recomendados e seus respectivos pesos associados. Este usuário possui apenas um trabalho, publicado em evento, que constituirá a fonte para geração de suas recomendações. A Tabela 4.3 apresenta os termos gerados e os respectivos pesos associados. Além disso, contém outras informações que são úteis para entendermos o cálculo de tais pesos como: ano, origem e idioma associados à publicação da qual os termos se originaram. O título "*Bibliographic Recommendations for individual users*" teve as *stopwords* removidas (no caso apenas a preposição "*for*") e o restante das palavras, cada uma delas, formou um termo simples ("*Bibliographic*", "*Recommendations*", "*individual*", "*users*"). Além disso, as palavras-chave "*Recommender Systems*" e "*Collaborative Filtering*", cada uma, constituiu um termo no vetor de consulta. Os seguintes valores de parâmetros foram adotados para o cálculo dos pesos auxiliares que compõe a Equação 4.1:

Para $w_{KeywordOrTitle}$: sendo i=1 para "palavras-chave" e i=2 para "título", n=2 (2 possibilidades: "título" ou "palavra-chave") e w_{min}=0,95 (valor utilizado neste exemplo).

Para $w_{Language}$: i=1 (proficiência de leitura do usuário no idioma Inglês é "bem"), n=3 (3 possibilidades: "bem", "razoavelmente" e "pouco") e w_{min}=0,60 (valor utilizado neste exemplo).

Para w_{Year}: i=2 (publicação do ano de 2005), n=4 (4 anos possíveis, do intervalo de 2006 a 2003) e w_{min}=0,55 (valor utilizado neste exemplo).

Dessa forma, os valores dos pesos auxiliares obtidos, de acordo com a Equação 4.2, são:

Se termo obtido de "título": $w_{KeywordOrTitle} = 1-(1-1).((1-0,95)/(2-1)) = 1,0$

Se termo obtido de "palavra-chave": $w_{KeywordOrTitle} = 1-(2-1).((1-0,95)/(2-1)) = 0,95$

$w_{Language} = 1-(1-1) .((1-0,60)/(3-1)) = 1,0$

$w_{Year} = 1-(2-1).((1-0,55)/(4-1)) = 0,85$

Assim, utilizando a Equação 4.1, para cálculo dos pesos de cada um dos seis termos da Tabela 4.3 (índices de 0 a 5), obtêm-se:

$w_0 = w_1 = w_2 = w_3 = (1,0).(0,95).(0,85) = 0,8075$

$w_4 = w_5 = (1,0).(1,0).(0,85) = 0,85$

Cabe salientar que, no caso do exemplo apresentado, os termos são originários de uma única publicação, mas eles podem aparecer em mais de uma publicação/curso de formação acadêmica do usuário, nestes casos, é utilizado pelo sistema o maior peso associado possível encontrado para o termo.

Tabela 4.3: Informações para montagem do vetor de consulta do *Usuário de Teste*

Índice Termo	*Termo*	*Peso*	*Idioma*	*Origem*	*Ano*
0	*BIBLIOGRAPHIC*	0,8075	Inglês	*Title*	2005
1	*RECOMMENDATIONS*	0,8075	Inglês	*Title*	2005
2	*INDIVIDUAL*	0,8075	Inglês	*Title*	2005
3	*USERS*	0,8075	Inglês	*Title*	2005
4	*RECOMMENDER SYSTEMS*	0,85	Inglês	*Keyword*	2005
5	*COLLABORATIVE FILTERING*	0,85	Inglês	*Keyword*	2005

Determinado o vetor de consulta, seguem-se os passos (3), (4) e (5) da Figura 4.13, da forma explicitada anteriormente, para gerar as recomendações. Ao final, no passo (6) são apresentadas as recomendações, sendo que um exemplo de um documento, possivelmente recomendado para o *Usuário Teste*, pode ser visto na Figura 4.15 (primeiro artigo recomendado pelo sistema a esse usuário). Nesta, encontram-se destacados os termos que compõem o vetor de consulta, os quais foram encontrados no título (*dc:title*) e resumo (*dc:description*) do artigo representado, publicado no XVI Simpósio Brasileiro de Inteligência Artificial, em 2002, indexado pela BDBComp.

Making Recommendations for Groups Using Collaborative Filtering and Fuzzy Majority; Sérgio R. de M. Queiroz; Francisco de A. T. de Carvalho; Geber L. Ramalho; Vincent Corruble; http://link.springer.de/link/service/series/0558/bibs/2507/25070248.htm; eng; 2002; XVI Simpósio Brasileiro de Inteligência Artificial.

In recent years, recommender systems have achieved a great success. Popular sites like Amazon.com and CDNow give thousands of recommendations every day. However, although many activities are carried out in groups, like going to the theater with friends, these systems are focused on recommending items for individual users. This brings out the need of systems capable of performing recommendations for groups of people, a domain that has received little attention in the literature. In this article we introduce an investigation of automatic group recommendations, making connections with problems considered in social choice and psychology. Then we suggest a novel method of making recommendations for groups, based on existing techniques of collaborative filtering and classification of alternatives using fuzzy majority. Finally we experimentally evaluate the proposed method to see its behavior under groups of different sizes and degrees of homogeneity.

Figura 4.15: Exemplo de um artigo recomendado ao *Usuário de Teste*

5 AVALIAÇÃO EXPERIMENTAL

Visando uma avaliação experimental do sistema de recomendação, foi solicitado a um grupo de indivíduos, formado por professores e alunos da Pós-Graduação do Instituto de Informática da UFRGS, vinculados aos grupos de pesquisa nas áreas de Sistemas de Informação, Banco de Dados e Computação Teórica, que disponibilizassem seus respectivos currículos Lattes. Nesta primeira avaliação, obtivemos um total de quatorze indivíduos para realização dos experimentos. Simultaneamente, a base de dados, contendo os artigos a serem recomendados, foi carregada através da colheita (*harvesting*) de metadados de todos os artigos cadastrados na BDBComp até junho de 2006, totalizando 3.978 artigos de 113 edições de conferências. Nestes experimentos, recomendações foram geradas pelo sistema, para cada indivíduo participante.

Duas avaliações foram desenvolvidas. A primeira baseava-se na hipótese de que os melhores artigos, para descrever o perfil do pesquisador, devem ser aqueles produzidos por ele próprio. Este tipo de análise, da recuperação dos artigos do próprio autor, é importante para avaliar a qualidade da recuperação implementada pelo sistema de recomendação. Levando em consideração que temos a informação sobre os artigos escritos por cada autor (pelo currículo), pode ser feito o "casamento" dos itens recomendados, ao pesquisador, que realmente foram escritos por ele. Esta avaliação foi realizada pelas métricas de revocação (*recall*) e precisão (*precision*) que são uma estratégia padrão de avaliação para sistemas de recuperação de informação (SALTON; MCGILL, 1983; BAEZA-YATES; RIBEIRO-NETO, 1999).

Revocação é usada para medir a porcentagem de documentos relevantes recuperados, em relação ao total que deveria ter sido recuperado. No caso da categorização de documentos, a métrica de revocação é usada para medir a porcentagem de documentos que foram corretamente classificados em relação ao número de documentos que deveria ter sido classificado.

Precisão é usada para medir a porcentagem de documentos corretamente recuperados, isto é, o número de documentos corretamente recuperados, dividido pelo número de documentos recuperados.

Como os perfis de usuários podem ser vistos como classes, e os artigos a serem recomendados como itens a serem classificados nesses perfis, pode-se verificar o total de itens do autor, que foram corretamente identificados (classificados) pelo perfil do usuário. Como existem diversos usuários (ou seja, muitas classes ou categorias), é necessário combinar os resultados. A macromédia (*macroaverage*), apresentada na Equação 5.1, foi desenvolvida por *D. Lewis* (LEWIS, 1991) para realizar a combinação especificada ("*the unweighted mean of effectiveness across all categories*"), e foi aplicada por ele na avaliação de algoritmos e técnicas de classificação.

$$macromédia = \frac{\sum_{i=1}^{n} X_i}{n} \quad (5.1)$$

Nesta equação, X_i é o valor de revocação ou precisão, dependendo da métrica que se deseja avaliar, para cada classe individual (usuário no nosso caso) e n é o número de classes (usuários). Assim, macromédia de revocação (*macroaverage recall*) é a média aritmética de revocações obtida para cada indivíduo e macromédia de precisão (*macroaverage precision*) é a média aritmética de precisões obtida para cada indivíduo.

Tendo em vista que cada usuário não está interessado em receber como recomendações seus próprios artigos, foi desenvolvida uma outra avaliação, que leva em consideração somente os itens de outros autores. Uma tela de exemplo da saída do sistema de recomendação, para um usuário participante do experimento, pode ser visualizada na Figura 5.1. Os artigos recomendados ao usuário são apresentados, ordenados decrescentemente, por um grau de relevância relativo gerado pelo próprio sistema. Neste *ranking*, o artigo com o maior grau de similaridade com o perfil do usuário recebe um percentual de 100% e os outros artigos têm seu percentual calculado em relação a este. Além do grau de relevância relativo, cada recomendação teve os seguintes atributos apresentados ao usuário: título do artigo (*dc:title*), autores (*dc:creator*), link para o artigo completo – URL (*dc:identifier*), idioma (*dc:language*), ano de publicação (*dc:date*), evento em que o mesmo foi publicado (*dc:source*) e resumo/abstract (*dc:description*).

Neste experimento, cada indivíduo foi convidado a avaliar as recomendações geradas para ele, atribuindo um dos cinco conceitos (segundo a escala bipolar de cinco pontos de Likert): “péssimo”, “ruim”, “médio”, “bom” ou “ótimo”, e também foi solicitado que os mesmos fizessem comentários sobre as recomendações recebidas. A seção seguinte apresenta os resultados obtidos.

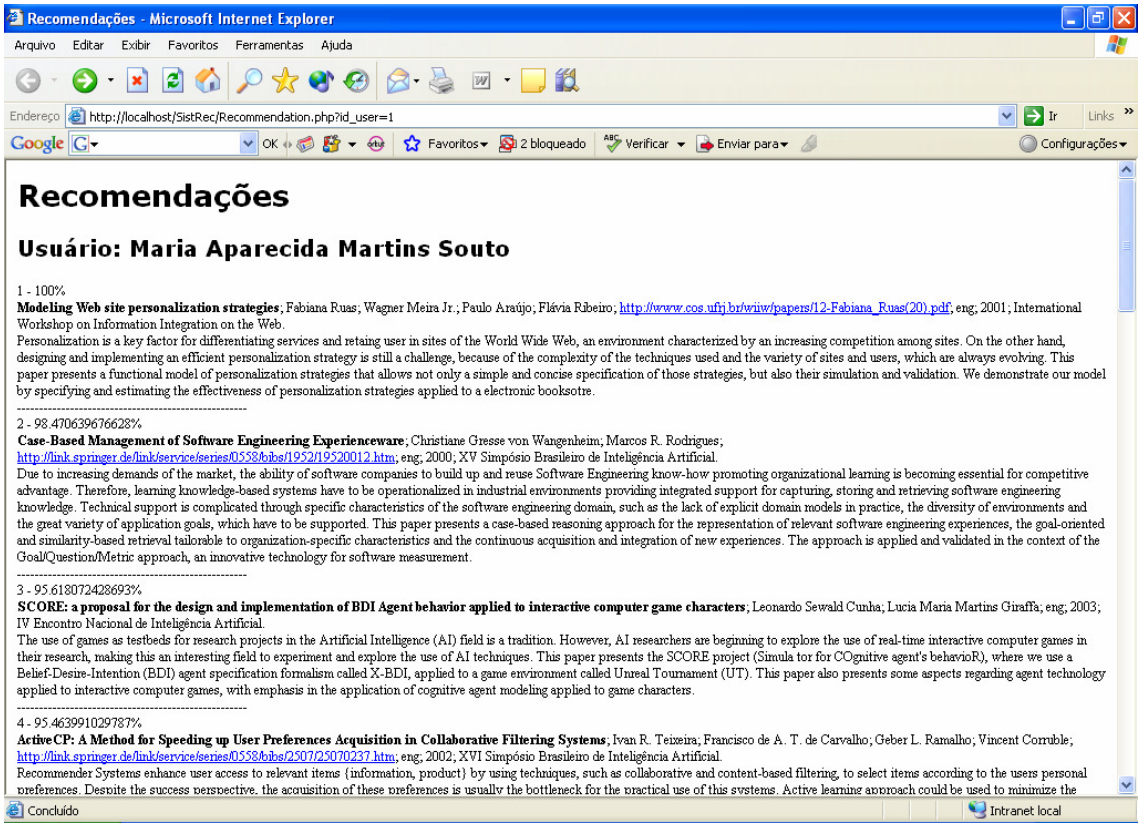

Figura 5.1: Tela exemplo de recomendações geradas

5.1 Análise dos experimentos

Nas seções seguintes, serão apresentados e discutidos os resultados obtidos através dos experimentos baseados em avaliações quantitativas e qualitativas das recomendações.

5.1.1 Avaliação Quantitativa

O primeiro experimento foi desenvolvido para avaliar a capacidade do sistema de, corretamente, identificar o perfil do usuário (isto é, para representar os interesses de pesquisa deste), sendo que, acreditamos serem os melhores artigos, para descrever o perfil do usuário, aqueles escritos por ele mesmo, como comentado anteriormente. Neste experimento, foram geradas 20 recomendações pelo sistema, para cada indivíduo participante. Para desenvolver tal avaliação, identificamos o número de artigos que cada autor possuía indexado na BDBComp. Depois disso, empregamos as métricas de revocação e precisão, para avaliar o número de artigos recuperados para cada autor de sua própria autoria e combinamos estas com a equação macromédia (Equação 5.1), explicada anteriormente.

Obtivemos o número de artigos que cada autor possuía indexado na BDBComp e concluímos que o valor máximo de macromédia de precisão que poderia ser obtido era de 23,18%. É importante salientar, que cada autor recebeu 20 recomendações, mas nenhum deles tinha 20 artigos de sua própria autoria indexados pela BDBComp (alguns, inclusive, não possuíam nenhum artigo indexado na mesma), o que explica o baixo valor de precisão obtido.

Neste primeiro experimento, duas situações distintas foram consideradas para montagem do perfil do usuário, levando em consideração: (i) período de informação de todo o currículo do usuário, e (ii) apenas as informações dos últimos três anos (incluindo o ano corrente, ou seja, 2003 a 2006) presentes em tais currículos.

Encontramos, neste experimento, utilizando apenas os últimos três anos de informação armazenadas no currículo Lattes, um percentual de macromédia de revocação de 43,25% e macromédia de precisão de 7,73%. Isto aconteceu porque o período de informações considerado é bem restrito. Sendo assim, artigos relacionados a áreas de interesse de pesquisa prévios do usuário não poderiam ser recomendados, já que o objetivo do sistema foi resumido à recomendação de artigos associados a áreas de interesse de pesquisa recentes do usuário.

Outra consideração importante a ser feita, é que o *ranking* das recomendações foi gerado com um grau de depreciação, dependente do ano de publicação e da proficiência de leitura do usuário nos idiomas, como explicado anteriormente (maiores detalhes ver seção 4.3, que explicita os valores dos parâmetros utilizados, nesta avaliação experimental, na atribuição de pesos ao vetor de consulta utilizado, para representar o perfil do usuário na geração das recomendações). Como o intervalo de tempo considerado corresponde a uma pequena parte do período completo de conferências armazenadas na BDBComp, nem todos os artigos constituiriam boas recomendações, já que o perfil do pesquisador pode se modificar ao longo do tempo.

Além disso, foi realizada uma outra análise, levando em consideração os valores obtidos caso todo o currículo dos pesquisadores fosse considerado, mas ainda há a depreciação dependente do ano das publicações presentes no currículo do usuário (ver seção 4.3 tratando dos valores de *i*, caso o intervalo de anos a serem considerados for omitido). Os resultados obtidos foram os seguintes: um percentual de macromédia de revocação de 63,25% e macromédia de precisão de 9,09%. Podemos observar que tais resultados são superiores aos descritos na situação anterior (ver Figura 5.2).

Dessa forma, verificamos que, à medida que uma maior quantidade de informações do perfil do usuário foi considerada, no processo de recomendação, um percentual maior de artigos do próprio usuário foram recomendados entre os 20 primeiros. Este comportamento já era esperado que fosse obtido pelo sistema. Além disso, cabe salientar que, como consideramos uma depreciação na importância das informações do perfil do usuário, de acordo com os anos de publicação, o foco principal (primeiros artigos a serem recomendados) ainda é a recomendação de artigos, que estejam relacionados com as áreas de interesse de pesquisa mais recentes do usuário. Diante desta situação, os resultados obtidos podem ser considerados bons.

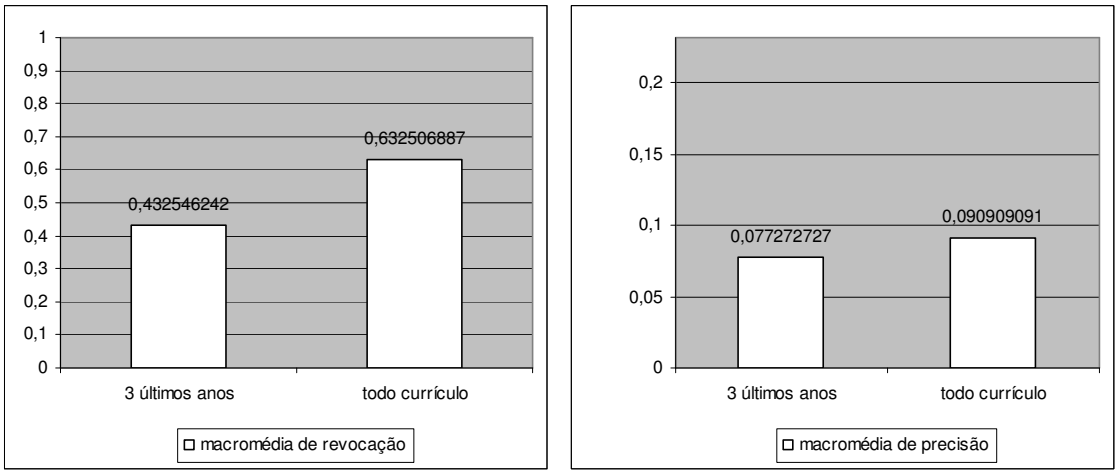

Figura 5.2: Avaliações Quantitativas (a) macromédia de revocação (b) macromédia de precisão

5.1.2 Avaliação Qualitativa

A próxima figura apresenta os resultados do segundo experimento, o qual baseia-se na avaliação qualitativa, dos artigos recomendados, realizada pelos usuários. Como explicado anteriormente, cada usuário recebeu 15 recomendações e avaliou-as, de acordo com um dos seguintes conceitos: "péssimo", "ruim", "médio", "bom" e "ótimo". Os resultados foram agrupados nas categorias "first match", "top 5", "top 10", e "top 15", e são apresentados na Figura 5.3.

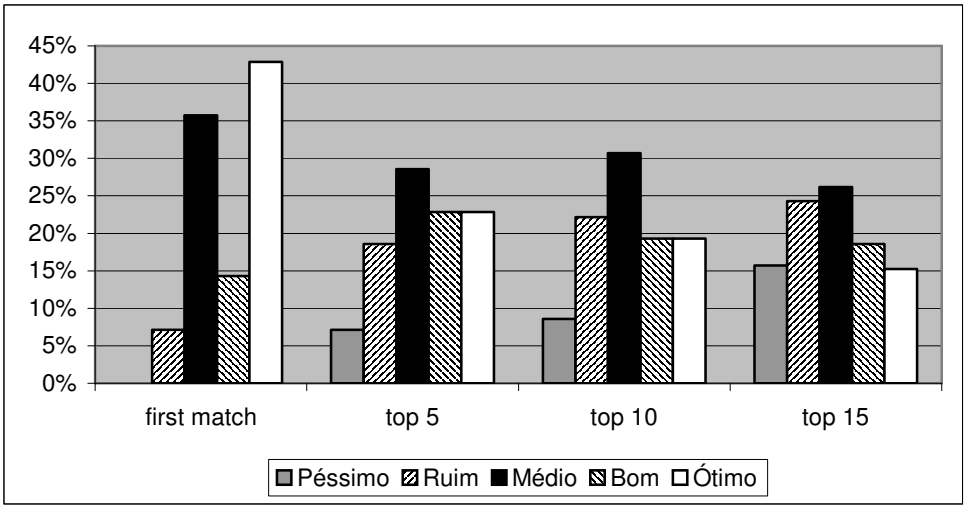

Figura 5.3: Avaliações das recomendações pelos usuários

Analisando estes resultados é possível observar que, se somente considerarmos o primeiro artigo recomendado ("first match"), o percentual de itens qualificados como "ótimo" é o maior de todos (42,86%) e que nenhum dos itens foi categorizado como "péssimo". Isto reforça a capacidade do sistema de gerar recomendações ajustadas aos interesses do usuário.

Agrupamos os conceitos "péssimo" e "ruim" em uma única categoria, referenciada como "recomendação negativa" e os conceitos "ótimo" e "bom" também em uma única categoria, referenciada como "recomendação positiva". Assim, podemos obter uma melhor visualização e compreensão dos resultados (Figura 5.4).

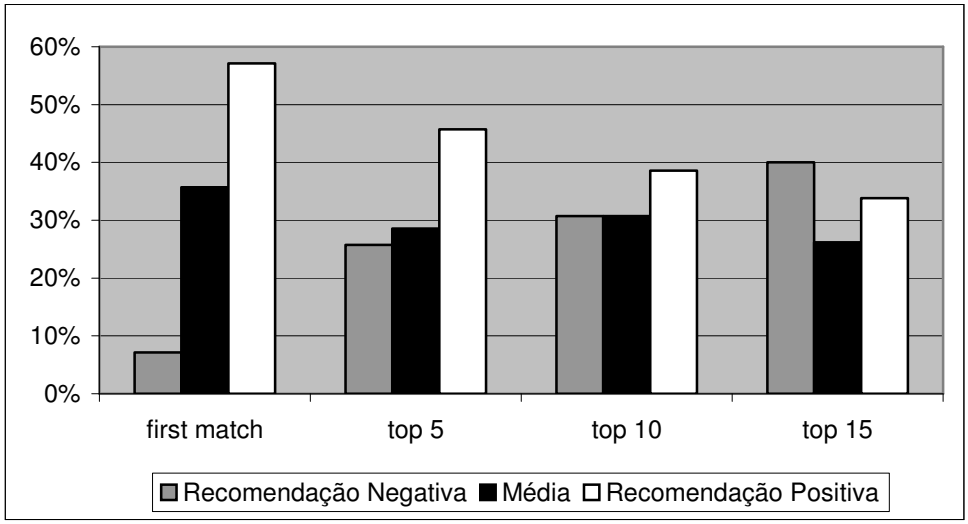

Figura 5.4: Avaliações dos usuários em categorias agrupadas

Podemos perceber que as recomendações positivas ("recomendação positiva"), considerando apenas o primeiro artigo recomendado ("first match"), são superiores (57,14%) em relação às recomendações negativas ("recomendação negativa") (apenas 7,14%). O mesmo comportamento pode ser percebido nas categorias "top 5" e "top 10", as recomendações tiveram uma avaliação negativa superior, apenas na categoria "top 15", e isto provavelmente aconteceu porque à medida que o número de recomendações cresce, o número de recomendações corretas diminui. Este é um comportamento esperado neste tipo de sistema e em sistemas de recuperação de informação em geral, onde, normalmente, conforme o valor de revocação do sistema cresce o valor de precisão cai. Ou seja, à medida que cresce o número de recomendações, decresce a sua qualidade (seu grau de similaridade com o perfil do usuário). Dessa forma, pode-se supor que o sistema de recomendação está ordenando adequadamente os artigos recomendados.

É importante observar o fato de que a BDBComp, atualmente, tem uma cobertura limitada, em virtude de não cobrir todas as áreas da Ciência da Computação, e isto pode ter influenciado negativamente a qualidade das recomendações. Fizemos esta observação com base em comentários feitos pelos próprios professores que participaram do experimento, como os seguintes:

> Acho que gerar tais resultados para mim pode ser complicado, pois na pós graduação trabalhei com duas áreas distintas, misturando temas de outras áreas ainda. Também, nas duas áreas em que mais tenho publicação, têm poucas pessoas trabalhando aqui no Brasil. Para resumir, diante de tais circunstâncias, a listagem que tu me enviaste está boa.
>
> Posso concluir que (a) não há muitos artigos na minha área ou (b) não estou sabendo descrevê-la corretamente […]

Assim, destaca-se como um trabalho futuro a necessidade de um maior estudo para estabelecimento de um limiar (*threshold*) de recomendação. A idéia é que artigos, com um grau de similaridade, em relação ao perfil do usuário, inferior ao valor de limiar estabelecido, não devem ser recomendados ao usuário.

Além disso, no experimento realizado, autores que mudaram de área de pesquisa nos últimos três anos, podem ter qualificado negativamente a recomendação de artigos que já lhes interessaram anteriormente. Outra possível causa de qualificações negativas é

que artigos, mesmo contendo várias palavras-chave utilizadas pelos autores para descreverem suas publicações no Lattes, podem não ter gerado uma boa recomendação se os contextos forem distintos. Mais ainda, as informações contidas no Lattes de alguns usuários podem não estar precisas o suficiente para gerarem recomendações consistentes. Existem alguns indícios a este respeito, conforme colocado por um dos usuários:

> [...] no meu caso, tive mais da metade das recomendações com grau 'péssimo'. Atribuo isso a 3 possíveis causas: 1) o abstract do material recomendado contém palavras-chave relacionadas aos meus trabalhos, mas o contexto na qual são aplicadas é totalmente diferente daquele que eu as utilizo [...] 2) as palavras-chave que uso no Lattes para caracterizar os meus artigos não estão precisas o suficiente; 3) eu tenho atuado em várias áreas, ou seja, não tenho me concentrado em um tema único.

Em uma outra análise, optamos por agrupar os usuários pesquisados em duas categorias: de professores e alunos da UFRGS. A Figura 5.5 apresenta as porcentagens de cada uma das qualificações agrupadas ("recomendação negativa", "média", "recomendação positiva"), por categoria de usuário, para as "top 15" recomendações.
Com esta divisão, pudemos observar que o sistema recebeu um percentual maior de qualificações positivas de professores do que de alunos. Esse também foi um dado interessante, que nos permitiu supor que os professores, pela sua experiência, conseguem descrever melhor seu interesse e possuem um maior número de publicações, o que nos leva a supor ainda, que o sistema obtém melhores resultados à medida que as informações no currículo Lattes são melhor descritas. Por outro lado, podemos também supor que os professores tenham um critério melhor definido e uma maior clareza, para avaliarem se as recomendações recebidas de fato atendem aos seus interesses e se estão relacionadas ao seu perfil como pesquisadores.

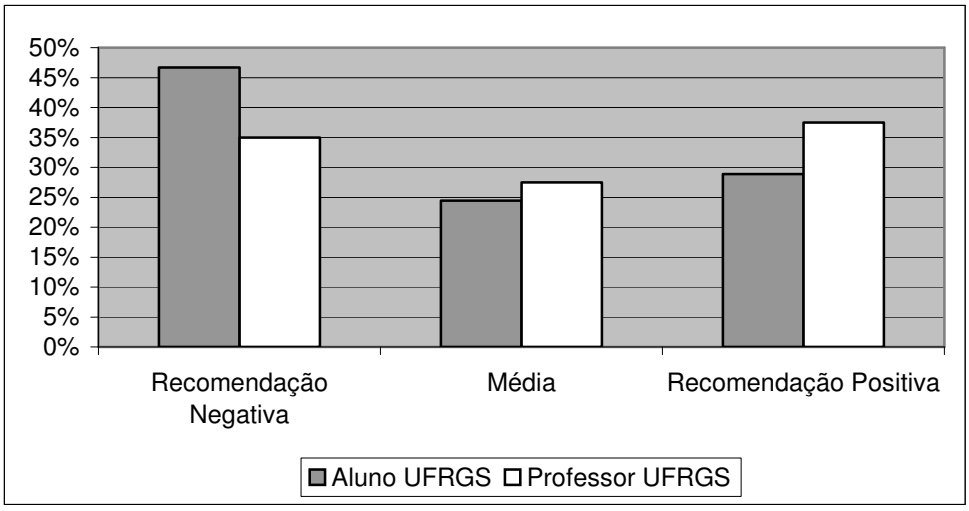

Figura 5.5: Resultados obtidos por categorias Aluno e Professor da UFRGS

Também foi realizada a análise de revocação *versus* precisão dos resultados obtidos previamente e consideramos como documentos relevantes aqueles classificados como "recomendação positiva".

A Figura 5.6 apresenta os pontos de precisão interpolados (normalizados) do sistema para os 11 níveis padrão de revocação. Este tipo de curva de precisão *versus* revocação é uma estratégia padrão para avaliação de sistemas de recuperação de informação, como sugerido por (BAEZA-YATES; RIBEIRO-NETO, 1999). Assim, para fazer o gráfico,

começamos localizando, para cada usuário, a posição de cada "recomendação positiva" (indicada pela avaliação realizada pelo usuário) na lista de 15 recomendações geradas pelo sistema. Para cada "recomendação positiva" calculamos os valores de revocação e precisão desta posição (ou seja, a porcentagem de documentos relevantes recuperados até este ponto). Como estes pontos podem estar em diferentes posições para cada usuário, é necessário normalizá-los para comparação e agregação. Para realizar tal normalização, para cada usuário foi preparada uma tabela, tendo os 11 pontos padrão de revocação (0; 0,1; 0,2; … ; 1), correspondendo a 0%; 10%; 20%; ... ; 100% de revocação. Então, mapeamos os valores encontrados para os pontos padrão seguindo a regra de interpolação descrita por (BAEZA-YATES; RIBEIRO-NETO, 1999), na qual cada ponto padrão é associado com o valor máximo de precisão conhecido, encontrado para qualquer nível de revocação maior ou igual ao nível que está sendo analisado.

Finalmente, para cada ponto padrão, calculamos uma média dos valores de precisão encontrados para cada usuário e geramos a curva apresentada na Figura 5.6.

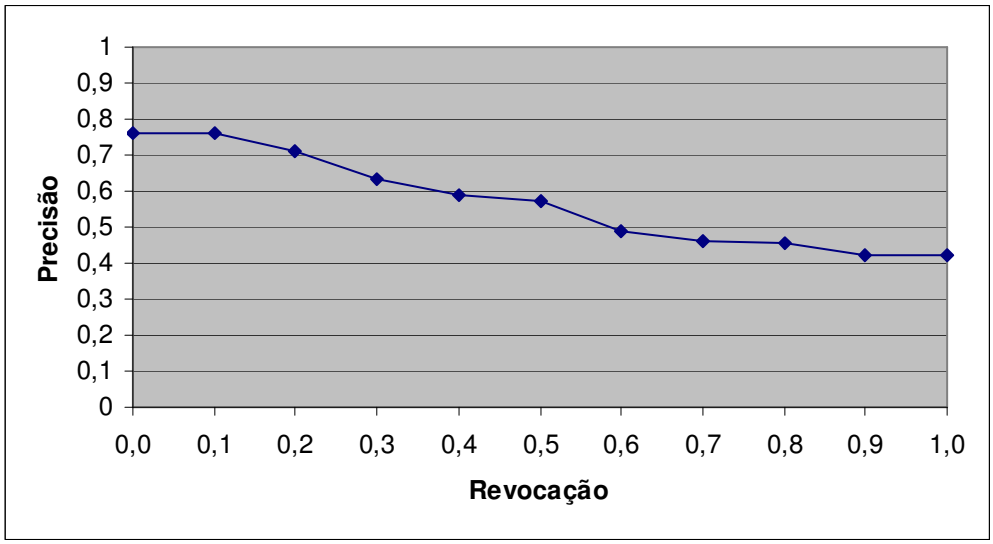

Figura 5.6: Precisão interpolada do sistema para 11 níveis padrão de revocação

Com esta curva interpolada e normalizada, é possível entender o comportamento global do sistema. No nosso experimento, foi possível observar que, para todos os usuários, para qualquer ponto de revocação (número de recomendações), a precisão do sistema foi acima de 40%, e para a primeira recomendação foi, em média, próximo de 76%. Este valor pode ser considerado bom, quando comparado aos obtidos por outras técnicas de recomendação tradicionais, nas quais, esse valor varia entre 60% e 90%.

Por fim, uma análise preliminar, comparando o sistema de recomendação proposto neste trabalho com uma outra abordagem, também foi desenvolvida. O sistema empregado na comparação também utiliza o modelo vetorial com a abordagem (tf x id) para a recuperação dos documentos a serem recomendados. Os termos obtidos do currículo Lattes do usuário, utilizados para a geração da consulta, também são os mesmos do sistema proposto, porém a abordagem para obtenção dos valores dos pesos associados a esses termos foi modificada. Os pesos de todos os termos receberam o valor 1, sendo considerados igualmente importantes para a consulta. Tal experimento foi realizado utilizando-se a mesma base de artigos a serem recomendados (artigos indexados pela BDBComp até junho de 2006). Dentre os 14 usuários que participaram dos experimentos, 10 avaliaram os cinco primeiros artigos recomendados a eles pelos

dois sistemas a serem comparados. Neste experimento, também levou-se em consideração apenas as informações dos últimos três anos do currículo dos usuários no processo de geração das recomendações. Os documentos considerados como relevantes são aqueles qualificados pelos usuários nas categorias “bom” ou “ótimo”.

Com base nas avaliações dos usuários, o gráfico de “Precisão interpolada do sistema para 11 níveis padrão de revocação”, apresentado na Figura 5.7, foi gerado para os resultados obtidos por cada um dos sistemas (da mesma forma como explicado anteriormente).

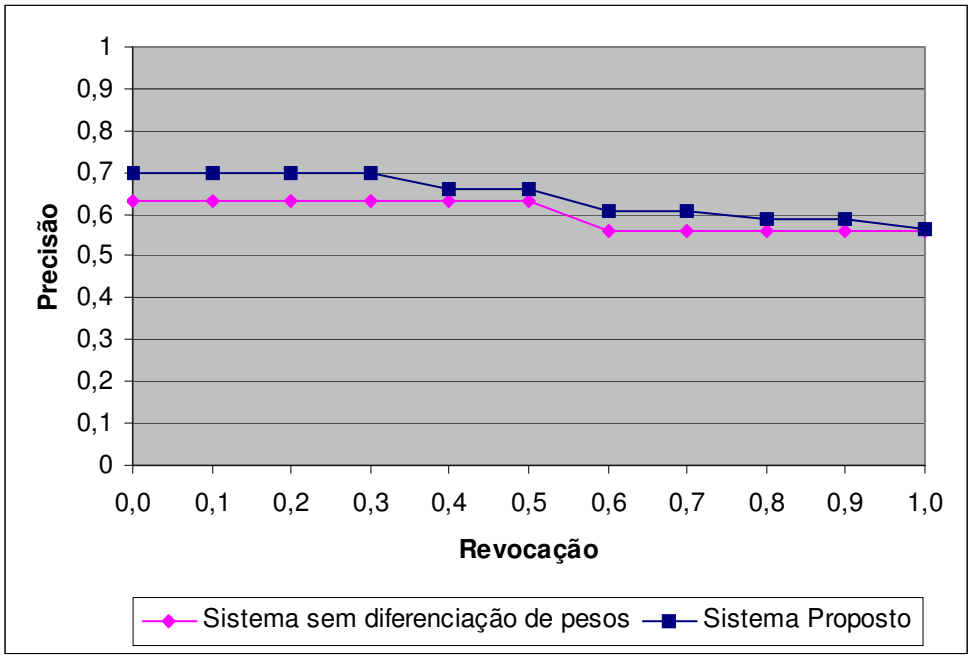

Figura 5.7: Precisão interpolada para 11 níveis padrão de revocação para dois sistemas distintos

Podemos observar que, os valores de precisão obtidos pelo nosso sistema de recomendação foram, em média, superiores ao da abordagem sem a diferenciação dos pesos associados aos termos da consulta, em todos os níveis de revocação. Para níveis de revocação de até 30% o valor de precisão da nossa abordagem ficou em torno de 70%, contra um valor de precisão em torno de 63% para outra abordagem considerada. Estes são resultados satisfatórios, que apresentam indícios importantes sobre o aumento na qualidade das recomendações geradas pelo nosso sistema e levantam a hipótese de que, a abordagem para atribuição dos pesos aos termos da consulta proposta neste trabalho, levando em consideração informações sobre o tipo do termo (obtido de “palavra-chave” ou “título”), o idioma e o ano de publicação, conduz a melhorias significativas na qualidade das recomendações geradas. Isto pode ser preliminarmente comprovado através do teste estatístico conhecido como *Student´s T-Test* (PRESS et al., 1999) o qual revelou que a diferença existente entre os resultados obtidos pelos dois sistemas é, realmente, estatisticamente significativa. Além disso, foi utilizado outro método que diz que uma diferença entre as precisões médias dos dois sistemas analisados sendo maior que 5% é notável e maior que 10% é material. Tal método, segundo Buckley & Voorhees (2000), é um método padrão utilizado na avaliação de sistemas de recuperação de informação, sendo utilizado em campanhas de avaliação como as do CLEF (*Cross Language Evaluation Forum*) e TREC (*Text REtrieval*

Conference). Dessa forma, a precisão média dos dois sistemas em questão foi calculada, sendo que a diferença obtida é um valor superior a 10%.

Este experimento de comparação apresentou resultados satisfatórios, mas precisa ser ampliado. Há a necessidade da obtenção de um maior volume de dados para a realização do experimento, incluindo um aumento no número de usuários e na quantidade de artigos recomendados avaliados. Outra observação importante é que, se forem considerados todos os anos do currículo do usuário, na montagem da consulta, este ganho tende a ser ainda maior, já que, considerar todas as informações com a mesma importância no currículo do usuário deve gerar a recomendação de artigos relacionados a interesses de pesquisa anteriores do usuário (que não necessariamente constituiriam boas recomendações atualmente) e ainda, tais artigos podem estar escritos em algum idioma que o usuário não possua qualquer proficiência de leitura. Entretanto, salientamos que esta é apenas uma análise preliminar e destacamos como trabalhos futuros a necessidade de um maior aprofundamento deste estudo.

6 CONCLUSÃO

O presente trabalho apresenta um Sistema de Recomendação para pesquisadores e acadêmicos da área da Ciência da Computação. Nos dias atuais, quando a descoberta de informação digital relevante na Web é uma tarefa complexa, tais sistemas são de grande valia para atenuar os problemas associados ao fenômeno da sobrecarga de informação, minimizando o tempo gasto para acessar as informações relevantes.

Uma característica importante deste trabalho consiste na utilização de informações padrão para a descrição tanto do perfil do usuário (currículo Lattes) quanto dos documentos de Bibliotecas Digitais (metadados no formato DC obtidos a partir de colheita) para gerar as recomendações. O sistema foi avaliado utilizando-se a BDBComp, mas pode ser utilizado para qualquer Biblioteca Digital que disponibilize os metadados de seus registros, utilizando o protocolo OAI-PMH. O mesmo ocorre com o *Curriculum Vitae* (CV) do usuário, o sistema, atualmente, trabalha com o currículo Lattes, mas pode trabalhar com qualquer arquivo XML que siga o subconjunto de informações do Lattes, utilizado para geração das recomendações neste trabalho. Este subconjunto pode ser obtido com informações advindas de outras fontes que não o sistema de currículos do CNPq, como por exemplo, Scholar Google, DBLP (*Digital Bibliography & Library Project*), páginas pessoais dos usuários, etc.

Um trabalho correlato importante é o provedor de dados compatível com o padrão OAI desenvolvido por (CONTESSA; OLIVEIRA, 2006) que determinou métodos para melhorias na quantidade e qualidade dos metadados disponibilizados sobre os trabalhos publicados por eventos da SBC (Sociedade Brasileira de Computação) no âmbito do sistema de submissões JEMS (*Journal and Event Management System*). A existência e a consistência dos valores de metadados, que descrevem os documentos digitais a serem recomendados, são importantíssimas para garantir a consistência das recomendações geradas. Esse provedor de dados, além de disponibilizar os artigos no formato Dublin Core Simples, também oferece a descrição no formato Dublin Core Qualificado. Dessa forma, podem ser estudadas extensões no modelo de recomendação proposto, para trabalhar também com o DC Qualificado, aproveitando-se dessa maior quantidade de informação disponibilizada na descrição dos documentos digitais, para possibilitar melhorias na qualidade das recomendações obtidas.

O sistema desenvolvido pode ter muitas aplicações. Uma delas é a recomendação de artigos para suportar o processo de aprendizagem, especialmente em sistemas de *eLearning*. A idéia básica é a seguinte: um estudante pode efetuar login num específico sistema de EAD (educação a distância) e receber recomendação de artigos, contendo material atualizado relevante ao seu perfil e que complemente o tópico de estudo corrente, visando à complementação de seus estudos.

O trabalho proposto foi desenvolvido no contexto dos projetos CTInfo, CNPq, DIGITEX, proc. 550845/2005-4 e PRONEX/FAPERGS, proc. 0408993, atualmente em andamento no Grupo de Modelagem Conceitual e Adaptabilidade da UFRGS. Nesse contexto, o trabalho desenvolvido pode ser utilizado no ambiente para Editoração, Indexação e Busca de Documentos Científicos em um Processo de Avaliação Aberta, que está sendo desenvolvido pelo grupo de pesquisa, cuja proposta foi apresentada em (OLIVEIRA et al., 2005). O objetivo da proposta apresentada é a criação de um ambiente onde todas as etapas da construção de um artigo científico sejam feitas de forma aberta, em especial as revisões. Segundo Oliveira et al. (2005), a proposta é que, imediatamente após ser submetido, cada artigo fique disponível à comunidade, a qual é formada por autores, revisores e comentaristas, os quais participam de discussões a respeito das temáticas abordadas por cada trabalho. Os artigos serão revisados, por revisores identificados, quando poderão surgir sugestões ou modificações e, paralelamente, os comentaristas poderão discutir o artigo em questão, podendo ainda, virem a surgir muitas contribuições.

A idéia é que, para o usuário pertencer ao ambiente ele envie seu currículo Lattes em formato XML. Pode ser disponibilizada uma descrição em DC dos artigos submetidos ao ambiente (formando a biblioteca digital) e o sistema de recomendação proposto poderá ser utilizado para sugerir artigos de leitura, que possivelmente irão interessar aos comentaristas, baseados em seus interesses de pesquisa. De forma similar, o sistema pode ser utilizado para indicar possíveis revisores aos trabalhos submetidos ao ambiente. O perfil do usuário, para as recomendações, pode ser enriquecido utilizando-se as informações referentes às suas interações com o ambiente em desenvolvimento (relacionadas aos trabalhos lidos, revisados e às respectivas avaliações dos outros usuários com relação às suas contribuições). Assim, um outro trabalho futuro pode ser a extensão do modelo, para trabalhar também com a abordagem de filtragem colaborativa, sendo utilizadas informações relativas aos interesses comuns dos usuários envolvidos no ambiente (trabalhos em conjunto ou revisões de trabalhos em áreas afins) para a geração das recomendações.

Além disso, podem ser realizadas extensões no modelo de recomendação proposto, para trabalhar com uma ontologia de domínio do usuário que está sendo desenvolvida como parte integrante dos projetos acima citados. Visa-se que tal ontologia possa enriquecer semanticamente a descrição do perfil do usuário e, por conseqüência, enriquecer as informações relevantes utilizadas no processo de recomendação. Um outro acréscimo importante ao trabalho desenvolvido é a consideração de uma ontologia das áreas de pesquisa, como a da ACM (*Association for Computing Machinery*), associadas aos documentos a serem recomendados e aos próprios pesquisadores, para auxiliar no processo de recomendação. Mas com relação a este ponto, o currículo Lattes possui sérias limitações, já que não possui uma taxonomia de áreas de pesquisa totalmente definida, sendo possível que o usuário informe esses dados de forma totalmente livre, ou mesmo não os informe.

O sistema de recomendação desenvolvido já permitiu uma experiência inicial satisfatória na recomendação de artigos publicados nos Simpósios Brasileiros da SBC. O grande interesse da comunidade local por este tipo de serviço ficou demonstrado pela alta taxa de resposta à avaliação e pelos comentários dos participantes. Isto nos estimula a disponibilizar, tão rápido quanto possível, uma versão Web do serviço no contexto do ambiente proposto em (OLIVEIRA et al., 2005).

REFERÊNCIAS

ADOMAVICIUS, G.; TUZHILIN, A. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. **IEEE Transactions on Knowledge and Data Engineering**, Los Alamitos, v.17, n.6, June 2005.

ARIADNE. **ARIADNE Project on Digital Libraries**. Disponível em: <http://www.comp.lancs.ac.uk/computing/research/cseg/projects/ariadne/>. Acesso em: jun. 2006.

ARP. **ARP - Active Recommendation Project**. Disponível em: <http://informatics.indiana.edu/rocha/lww/>. Acesso em: jun. 2006.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. Wokingham, UK: Addison-Wesley, 1999.

BALABANOVIC, M.; SHOHAM, Y. Combining Content-Based and Collaborative Recommendation. **Communications of the ACM**, New York, v.40, n.3, Mar. 1997.

BDBComp. **Biblioteca Digital Brasileira de Computação**. Disponível em: <http://www.lbd.dcc.ufmg.br/bdbcomp/>. Acesso em: out. 2005.

BERNERS-LEE, T. **Weaving the Web:** the original design and ultimate destiny of the World Wide Web. San Francisco: HarperCollins, 1999.

BUCKLEY, C.; VOORHEES, E. M. Evaluating evaluation measure stability. In: ANNUAL INTERNATIONAL ACM/SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 23., 2000, Athens, Greece. **Proceedings...** New York: ACM Press, 2000. p.33-40.

CALLAN, J. et al. **Personalisation and Recommender Systems in Digital Libraries**. 2003. Disponível em: <http://www.dli2.nsf.gov/internationalprojects/working_group_reports/personalisation.pdf> Acesso em: dez. 2005.

CAZELLA, S. C.; REATEGUI, E. **Recommender Systems**. Minicurso apresentado no ENIA, 2005. Disponível em: <http://www.inf.unisinos.br/%7Ecazella/papers/enia2005.pdf >. Acesso em: dez. 2005.

CITIDEL. Disponível em: <http://www.citidel.org/>. Acesso em: nov. 2005.

CLAYPOOL, M. et al. Combining content-based and collaborative filters in an online newspaper. In: ACM/SIGIR WORKSHOP ON RECOMMENDER SYSTEMS: ALGORITHMS AND EVALUATION, 1999, Berkeley, California. **Proceedings…** New York: ACM Press, 1999.

CLEF and Multilingual information retrieval. [S.l.]: Institut interfacultaire d'informatique, University of Neuchatel. Disponível em: <http://www.unine.ch/info/clef/>. Acesso em: dez. 2005.

CONTESSA, D. F.; OLIVEIRA, J. P. M. de. An OAI Data Provider for JEMS. In: ACM SYMPOSIUM ON DOCUMENT ENGINEERING, 6., 2066, Amsterdam. **Proceedings...** New York: ACM, 2006. p.218 - 220.

COSLEY, D.; LAWRENCE, S.; PENNOCK, D. M. REFEREE: an open framework for practical testing of recommender systems using ResearchIndex. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, VLDB, 28., 2002. **Proceedings...** [S.l.: s.n.], 2002. p.35-46.

CyberStacks. Disponível em: <http://www.public.iastate.edu/~CYBERSTACKS/>. Acesso em: jun. 2006.

DCMI. **History of the Dublin Core Metadata Initiative**. Disponível em: <http://dublincore.org/about/history/>. Acesso em: out. 2006.

DC-OAI. **A XML schema for validating Unqualified Dublin Core metadata associated with the reserved oai_dc metadataPrefix**. Disponível em: <http://www.openarchives.org/OAI/2.0/oai_dc.xsd>. Acesso em: out. 2005.

DOLOG, P.; NEJDL, W. **Personalisation in Elena:** How to cope with personalisation in distributed eLearning Networks. 2003. Disponível em: <http://www.l3s.de/~dolog/pub/sinn2003.pdf>. Acesso em: set. 2005.

DUBLIN Core Metadata Initiative. Disponível em: <http://dublincore.org>. Acesso em: set. 2005.

GROSSMAN, D. A. **Information retrieval:** algorithms and heuristics. 2nd ed. Dordrecht: Springer, c2004. 332p.

GUTTERIDGE, C. **GNU EPrints 2 overview**. [S.l.: s.n.], 2002.

HERLOCKER, J. L. **Understanding and Improving Automated Collaborative Filtering Systems**. 2000. 140 f. Tese (Doutorado em Ciência da Computação), University of Minnesota, Minnesota.

HERLOCKER, J. L.; KONSTAN, J. A. Content-Independent Task-Focused Personalization and Privacy Recommendation. **IEEE Internet Computing**, New York, v.5, n.6, p.40-47, 2001.

HILLMANN, D. **Using Dublin Core**. Nov. 2005. Disponível em: <http://dublincore.org/documents/usageguide/>. Acesso em: dez. 2005.

HUANG, Z. et al. A Graph-based Recommender System for Digital Library. In JCDL, 2002. **Proceedings...** [S.l.: s.n.], 2002.

HWANG, S. Y.; HSIUNG, W. C.; YANG, W. S. A prototype WWW literature recommendation System for Digital Libraries. **Online Information Review**, [S.l.], v.27, n.3, p.169-182, 2003.

JONES, K. S.; WALKER, S.; ROBERTSON, S. E. **A probabilistic model of information retrieval:** development and status. Cambridge: Cambridge University Computer Laboratory, 1998. (TR 446).

LAENDER, A. H. F.; GONÇALVES, M. A.; ROBERTO, P. A. BDBComp: Building a Digital Library for the Brazilian Computer Science Community. In: JOINT CONFERENCE ON DIGITAL LIBRARIES, JCDL, 4., 2004, Tucson, Arizona. **Proceedings...** New York: ACM, 2004. p.23-24.

LAGOZE, C.; SOMPEL, H. V. de. The Open Archives Initiative: Building a low-barrier interoperability framework. In: JOINT CONFERENCE ON DIGITAL LIBRARIES, JCDL, 2001, Roanoke, Virginia. **Proceedings...** New York: ACM, 2001.

Lattes-CNPq: Plataforma Lattes - Conselho Nacional de Desenvolvimento Científico e Tecnológico. Disponível em: <http://lattes.cnpq.br/>. Acesso em: out. 2005.

LEWIS, D. D. Evaluating text categorization. In: SPEECH AND NATURAL LANGUAGE WORKSHOP, 1991, Pacific Grove, California. **Proceedings...** Morristown, New Jersey: Association for Computational Linguistics, 1991. p.312-318.

LPML-CNPq. **Padronização XML:** Curriculum Vitae. Disponível em: <http://lmpl.cnpq.br/lmpl/?go=cv.jsp>. Acesso em: out. 2005.

MALY, K. et al. Light-weight communal digital libraries. In: JOINT CONFERENCE ON DIGITAL LIBRARIES, JCDL, 4., 2004, Tucson, Arizona. **Proceedings…** New York: ACM, 2004. p.237-238.

MySQL: The world's most popular open source database. Disponível em: <http://www.mysql.com/>. Acesso em: mar. 2006.

NIEDERAUER, J. **PHP com XML – Guia de Consulta Rápida**. São Paulo: Novatec, 2002. 96 p.

OAI Repository Explorer: Open Archives Initiative - Repository Explorer. Disponível em: <http://re.cs.uct.ac.za/>. Acesso em: dez. 2006.

OAI: Open Archives Initiative. Disponível em: <http://openarchives.org>. Acesso em: out. 2005.

OAI-PMH: The Open Archives Initiative Protocol for Metadata Harvesting. Disponível em: <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>. Acesso em: nov. 2005.

OLIVEIRA, J. P. M. de; GALANTE, R. de M.; MUSA, D. L.; EDELWEISS, N. Uma Proposta para Editoração, Indexação e Busca de Documentos Científicos em um Processo de Avaliação Aberta. In: WORKSHOP EM BIBLIOTECAS DIGITAIS, WDL, SBBD/SBES, 1., 2005. **Anais...** Rio de Janeiro: SBC, 2005. p.30-39.

PHP Group. **PHP: Hypertext Preprocessor**. Disponível em: <http://www.php.net/>. Acesso em: mar. 2006.

PRESS, W. A. et al. **Numerical Recipes in C:** The Art of Scientific Computing. 2nd ed. New York: Cambridge University Press, 1992. 616 p.

RENDA, M. E.; STRACCIA, U. A Personalized Collaborative Digital Library Environment: a model and an application. In: INTERNATIONAL CONFERENCE ON ASIAN DIGITAL LIBRARIES, ICADL, 5., 2002, Singapore, Republic of Singapore. **Proceedings...** [S.l.]: Springer-Verlag, 2002. p.262-274.

ROCHA, L. M. TalkMine: A Soft Computing Approach to Adaptive Knowledge Recommendation. In: LOIA, V.; SESSA; S. (Ed.) **Soft Computing Agents:** New Trends for Designing Autonomous Systems. [S.l.]: Physica-Verlag, 2001.

SALTON, G.; BUCKLEY, C. Term-Weighting Approaches in Automatic Text Retrieval. **Information Processing and Management**, [S.l.], v.24, n.5, p.513-523, 1988.

SALTON, G.; FOX, E. A.; WU, H. Extended Boolean information retrieval. **Communications of the ACM**, New York, v.26, n.11, p.1022-1036, Nov. 1983.

SALTON, G.; MCGILL, M. J. **Introduction to Modern Information Retrieval**. New York: McGraw-Hill Book. 1983. 448 p.

SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. **Communications of the ACM**, New York,v.18, n.11, p.613-620, Nov. 1975.

SHAHABI, C.; CHEN, Y. S. An Adaptive Recommendation System without Explicit Acquisition of User Relevance Feedback. **Distributed and Parallel Databases**, [S.l.], v.14, n.2, p.173-192, Sept. 2003.

SILVA, L. V. e; LAENDER, A. H. F.; GONÇALVES, M. A. **The BDBComp Self-Archiving Service:** Design Issues and Usability Evaluation. Disponível em: <http://i3dl.lbd.dcc.ufmg.br/files/BDBCompSelfArch.pdf>. Acesso em: dez. 2005.

SOMPEL, H. V. de; LAGOZE, C. The Santa Fe Convention of the Open Archives Initiative. **D-Lib Magazine**, [S.l.], v.6, n.2, Feb. 2000.

TANSLEY, R. et al. DSpace: An institutional digital repository system. In: JOINT CONFERENCE ON DIGITAL LIBRARIES, JCDL, 3., 2003, Houston, Texas. **Proceedings…** New York: ACM, 2003.