Sistema de Recomendação para Bibliotecas Digitais sob a Perspectiva da Web Semântica[1]

Giseli Rabello Lopes, Maria Aparecida Martins Souto, José Palazzo Moreira de Oliveira

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil

{grlopes,souto,palazzo}@inf.ufrgs.br

Abstract. *This paper describes a Recommendation System of scientific articles in digital libraries. In particular the system is planned for the scientific community of Computer Science area. Technologically, the proposed system was developed under Semantic Web perspective, once it explores their emergent technologies as: standard metadata for describing the documents - Dublin Core, use of XML standard for describing the user profile - Lattes Curriculum Vitae, and use of service and data providers to generate the recommendation. In addition, this paper presents and discusses some results of the experimental evaluation.*

Resumo. *Este artigo 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. 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, uso de padrão XML para a descrição do perfil do usuário - Currículo Lattes, e utilização de provedor de serviços e dados para gerar a recomendação. Este artigo ainda apresenta e discute os resultados da avaliação experimental realizada.*

1. Introdução

Atualmente, pesquisadores e acadêmicos têm se beneficiado 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 (BDs) 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] e Kepler [Maly et al. 2004]. No Brasil deve ser citada a

[1] Este trabalho é parcialmente financiado pelo projeto CTInfo, CNPq, DIGITEX, proc. 550845/2005-4 e pelo projeto PRONEX/FAPERGS, proc. 0408993. A primeira autora é apoiada por bolsa de mestrado do CNPq, o terceiro autor é parcialmente apoiado por bolsa de produtividade do CNPq.

BDBComp (Biblioteca Digital Brasileira de Computação) [Laender et al. 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 interesses são igualmente providos com a mesma informação, em resposta a uma mesma consulta. Com o objetivo de suprir estas dificuldades, Sistemas de Recomendação para BDs têm sido desenvolvidos, entre os quais citamos os projetos ARIADNE, ResearchIndex, CyberStacks e ARP.

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]. Esta funcionalidade pode ser obtida através da adaptação da apresentação, conteúdo e/ou serviços baseados na atividade da pessoa, bagagem cognitiva, histórico, necessidades de informação, localidade, etc.

Este estudo insere-se no contexto acima exposto. Especificamente, este artigo apresenta um Sistema de Recomendação de artigos científicos, na área da Ciência da Computação, que estejam de acordo com os interesses do usuário, os quais são 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 Core 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.

O artigo está organizado da seguinte maneira: a seção 2 apresenta o contexto tecnológico no qual o sistema de recomendação foi desenvolvido. A seção 3 detalha a arquitetura e o módulo de recomendação do sistema proposto. As seções 4 e 5 descrevem o experimento de validação do sistema e apresenta alguns resultados importantes obtidos. Por fim, a seção 6 apresenta algumas considerações sobre o sistema e os resultados da avaliação realizada, bem como apresenta os trabalhos futuros.

2. Sistemas de Recomendação em BDs sob a perspectiva da Web Semântica

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* apresenta duas grandes vantagens: obtenção de maior eficiência computacional durante a colheita de informações a serem recomendadas; e possibilidade de se obter *interoperabilidade* entre as BDs. 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 Core 2005].

A *Open Archives Initiative* [OAI 2005] teve um papel muito importante para permitir a interoperabilidade entre as BDs. Seu principal objetivo foi o de fazer com que diferentes BDs ao redor do mundo pudessem interoperar formando uma federação [Sompel et al. 2000]. A partir de então, ficou definida uma forma padrão de comunica-

ção entre BDs. Assim, a colheita de metadados (*harvesting*) por parte das BDs é feita utilizando-se o protocolo OAI-PMH (*Open Archives Initiative Protocol for Metadata Harvesting*) [OAI-PMH 2005] que define como deve ser realizada a transferência de metadados entre duas entidades básicas: provedores de dados e de serviços. Os provedores de dados têm a função de buscar metadados em bases de dados e disponibilizá-los suportando o protocolo OAI-PMH. Os provedores de serviços utilizam os metadados disponibilizados pelo provedor de dados para fornecer serviços mais específicos. A interação entre as duas entidades básicas do OAI-PMH pode ser vista na Figura 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 coletados, o provedor de serviços pode, então, oferecer um determinado serviço como, por exemplo, um sistema de busca ou recomendação.

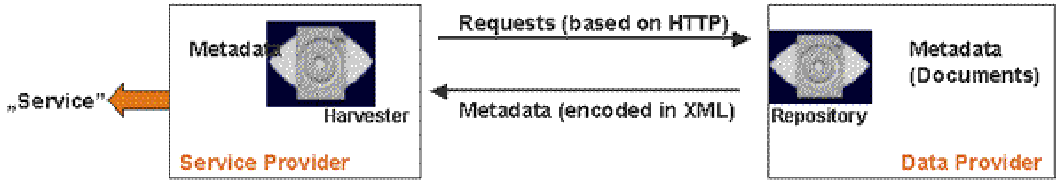

Figura 1. Interação entre as entidades básicas do OAI-PMH [OAI 2005]

Para que seja possível a tarefa de *harvesting* dos metadados de provedores de dados são definidos seis tipos de requisições chamadas de "verbos" que são: *Identify*, *ListMetadataFormats*, *ListSets*, *ListIdentifiers*, *ListRecords*, e *GetRecords*. Para maiores detalhes ver [OAI-PMH 2005].

Outro ponto importante para garantir a interoperabilidade é a adoção de um padrão básico para descrição dos metadados pelos provedores de dados. Para tanto, foi escolhido o formato Dublin Core por este ser um formato simples que permite a descrição dos recursos disponíveis através de um conjunto mínimo de metadados e que pode ser codificado em XML. O formato Dublin Core simples é formado por 15 elementos (*Dublin Core Metadata Element Set*) que são: *title*, *creator*, *subject*, *description*, *publisher*, *contributor*, *date*, *type*, *format*, *identifier*, *source*, *language*, *relation*, *coverage* e *rights*. A descrição completa de tal conjunto de elementos pode ser obtida em [Dublin Core 2005].

Em relação aos sistemas de recomendação, sob o ponto de vista metodológico, existem três tipos básicos de abordagens utilizadas [Huang et al. 2002]: (i) *filtragem baseada em conteúdo*, que recomenda itens classificados no perfil do usuário, e que são similares a outros pelos quais o usuário já demonstrou interesse. Neste tipo de abordagem, não existe surpresa na recomendação já que itens que não se relacionam com o perfil do usuário não serão recomendados; (ii) *filtragem colaborativa*, que utiliza informações de outros usuários, trabalhando com a idéia de que se os interesses dos usuários são similares, itens preferidos por um usuário podem ser recomendados a outros. Nesse tipo de abordagem podem ocorrer problemas como a "partida fria" (*coldstart*) quando não estão inicialmente disponíveis dados sobre o perfil do usuário; e (iii) *híbrida*, que é uma combinação das duas abordagens apresentadas anteriormente buscando agregar as características de cada uma delas e solucionar as limitações encontradas. O sistema de

recomendação aqui apresentado utiliza a abordagem baseada em conteúdo. Isto porque a idéia, neste caso, é combinar apenas as informações do usuário obtidas a partir do seu currículo Lattes com as informações referentes aos artigos para gerar a recomendação personalizada.

Para o processo de recuperação de informação diversos modelos podem ser adotados, como os clássicos: booleano, vetorial ou probabilístico, ou ainda, modelos como o *fuzzy* e o booleano estendido [Ferneda 2003]. Dentre os modelos de indexação existentes é adotado, neste trabalho, o modelo vetorial [Salton et al. 1988], o qual representa documentos e consultas como vetores de termos e devolve como resultado do processo de recomendação, documentos ordenados através de um cálculo de similaridade.

Em tal modelo, uma expressão de busca, ou seja, a consulta que se deseja realizar para retornar documentos relevantes é representada com um vetor de termos que é um vetor de palavras ou expressões, as quais têm associadas a si um grau de relevância (peso) para consulta. Este peso representa a coordenada do vetor de busca em tal dimensão. De forma análoga, para se calcular a similaridade entre os documentos e uma expressão de busca, o documento também é representado por um vetor de termos, sendo atribuído a cada termo um peso que indica a importância do respectivo termo na descrição do documento.

O princípio do modelo vetorial é que vetores localizados espacialmente mais próximos representam documentos mais similares. Dessa forma, para o cálculo do valor de similaridade (*Similarity*) de um documento (D) em relação a uma expressão de busca (Q) é utilizada a fórmula dada abaixo, onde "w" representa os pesos e "t" o número total de documentos considerados. A fórmula utilizada é a do co-seno (originada do produto escalar entres dois vetores) formado entre o vetor de busca desejado e o vetor do documento. O valor de similaridade encontrado indicará o grau de relevância do documento em resposta à expressão de busca. Para maiores detalhes ver [Salton et al. 1988].

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

Pode-se observar que tal sistema de indexação de textos é baseado na atribuição de pesos apropriados aos termos dos vetores; dessa forma, o resultado obtido depende crucialmente da escolha de sistemas de atribuição de pesos que sejam efetivos. Além disso, a determinação dos termos que comporão a expressão de busca é fundamental para se obter um resultado de recomendação adequado às necessidades do usuário. Com relação a estas questões, serão apresentadas, na seção 3.2 deste artigo, as decisões adotadas para o sistema de recomendação proposto neste trabalho.

3. O sistema de recomendação

A seguir apresentamos a arquitetura do Sistema de Recomendação, assim como detalhamos o módulo de recomendação propriamente dito.

3.1. Arquitetura do Sistema

O trabalho descrito no presente artigo propõe um Sistema de Recomendação em Bibli-

otecas Digitais, sob a perspectiva da Web Semântica [Berners-Lee 1999], ou seja, propõe-se a implementar uma máquina de busca enriquecida semanticamente através do uso de padrões de metadados que descrevem não somente a forma sintática dos documentos e seus conteúdos semânticos, mas também descrevem características de pessoas, potencialmente usuárias destes documentos. 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 Dublin Core codificados em XML. O padrão XML para o *Curriculum Vitae* da Plataforma Lattes é mantido pela Comunidade CONSCIENTIAS-LMPL. 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 esquema do XML utilizado como fonte de dados dos documentos digitais que é obtido como resultado do processo de colheita em [DC-OAI 2005]. O sistema consiste de um provedor de serviços cuja arquitetura é apresentada na Figura 2.

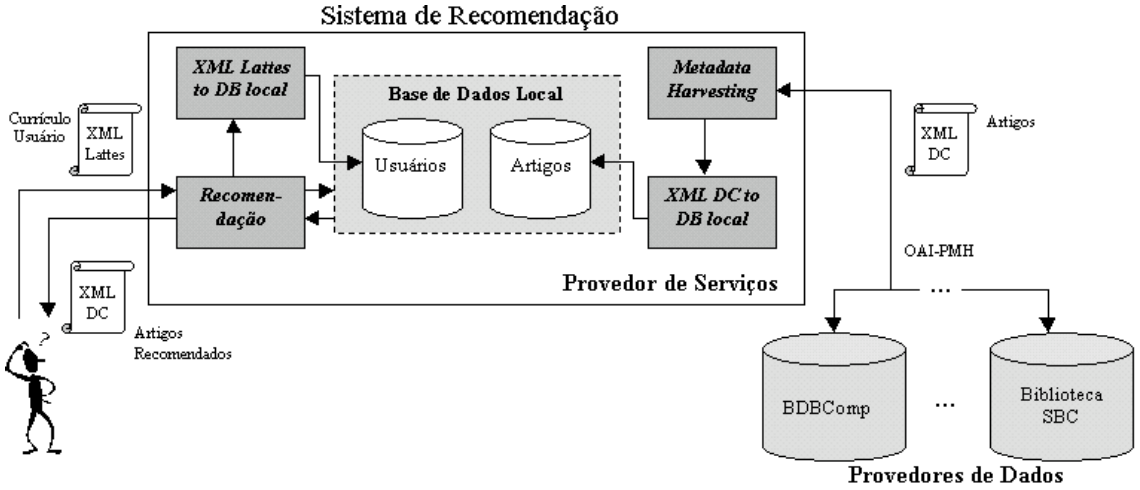

Figura 2. Arquitetura do Sistema

O funcionamento do sistema será descrito a seguir. Para armazenamento das informações dos documentos digitais a serem recomendados: (i) É ativado o módulo ***Metadata Harvesting*** para fazer a colheita de metadados no formato Dublin Core (DC) em XML que descrevam documentos digitais armazenados em determinada Biblioteca Digital, sendo para tanto utilizado o protocolo OAI-PMH. O foco do sistema de recomendação proposto é a recomendação de artigos na área da Ciência da Computação publicados em congressos; (ii) Com base nos arquivos XML (contendo metadados de artigos de determinado congresso descritos no formato DC) coletados pelo módulo anteriormente descrito é ativado o módulo ***XML DC to DB local*** que irá interpretar o(s) XML(s) recebido(s) e armazenar as informações relevantes para a recomendação na base de dados local, além é claro, da catalogação do próprio arquivo XML recebido.

Após os artigos terem sido armazenados, pode-se efetuar a recomendação destes, sendo que, para tanto, é ativado o módulo ***Recomendação*** no qual: (i) O usuário que deseja receber uma recomendação se cadastra no sistema e envia seu currículo Lattes no formato XML; (ii) O módulo ***XML Lattes to DB local*** é ativado, sendo as informações relevantes para a tarefa de recomendação contidas em tal XML armazenadas na base de dados local; (iii) Com base nas informações armazenadas na base local referentes ao

currículo do usuário e também daquelas referentes aos artigos, o módulo de ***Recomendação*** realiza a tarefa de recomendação propriamente dita; (iv) Como resultado da recomendação é gerado um XML (no formato DC) contendo os artigos recomendados a determinado usuário. A seguir será descrito em detalhes o serviço de recomendação propriamente dito, implementado pelo provedor de serviços.

3.2. O módulo de recomendação

Para determinar os termos do vetor de busca são utilizadas as informações "título" e "palavras-chave" contidas no Lattes sobre a formação acadêmica (graduação, mestrado ou doutorado) e a produção bibliográfica do usuário (trabalho em eventos, artigos publicados, texto em jornal ou revista, livro publicado ou organizado, capítulo de livro publicado). Para composição do vetor de busca, das palavras contidas em "título" são eliminadas as *stopwords* [CLEF 2005] (termos cujo conteúdo semântico é limitado, sendo muito freqüentes em todos os documentos, como preposições, artigos e conjunções). Assim, cada palavra restante do "título" será um termo simples. As palavras contidas em "palavras-chave" são adicionadas integralmente formando um termo composto. Visando reduzir o espaço de busca, a recomendação considera apenas os idiomas que o usuário tem proficiência de leitura. Além disso, publicações mais recentes e cursos de formação acadêmica em andamento também terão preferência.

Dessa forma, a atribuição dos pesos dos termos de busca foi realizada de acordo com a combinação das seguintes características: (i) *aos termos obtidos de "palavras-chave" foram atribuídos pesos maiores do que a termos obtidos do "título"*, 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; (ii) *pesos dos termos de acordo com o idioma*, atribuído de acordo com a proficiência de leitura do usuário no idioma do termo, peso em ordem crescente para "pouco", "razoavelmente" e "bem"; (iii) *peso maior para termos obtidos de cursos de formação acadêmica e publicações mais recentes*, ou seja, 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.

Após determinado o vetor de busca, é necessário calcular os pesos dos termos que compõe a expressão de busca em cada um dos documentos que possivelmente serão recomendados, formando assim os vetores dos documentos. A abordagem adotada nesse trabalho é a (tf X idf) (*the product of the term frenquency and the inverse document frequency*), apresentada em [Salton et al. 1988], que é utilizada para atribuição automática de pesos aos termos para recuperação de textos. Na abordagem (tf X idf), duas medidas são utilizadas em conjunto: (i) *term frequency (tf)* é a freqüência de um termo, sendo definida pelo número de vezes que determinado termo aparece no documento; (ii) *inverse document frequency (idf)* é o inverso da freqüência do termo na coleção, este fator varia inversamente com o número de documentos "n" nos quais um termo aparece de uma coleção de "N" documentos, tipicamente calculado como log (N/n). A idéia de Salton sugere que os melhores termos para identificação do conteúdo do documento são aqueles capazes de distinguir certos documentos individuais do restante da coleção. Isso implica que os melhores termos de indexação, ou seja, os que apresentarão maior peso, devem ter valores altos de *tf* e *idf*. Como no caso do sistema proposto estão sendo utilizadas informações contidas em metadados do formato Dublin Core que descrevem os

artigos, os elementos *dc:title* e *dc:description* serão utilizados como representantes do conteúdo do artigo em si, para o cálculo das métricas apresentadas. Além disso, como o sistema pode lidar com idiomas distintos, o número total de documentos da coleção irá variar de acordo com o idioma do termo considerado.

Determinados os termos dos vetores de busca e de documentos e calculados seus respectivos pesos, poderão ser obtidos os valores de similaridade de cada documento da base de dados em relação à busca desejada, de acordo com a fórmula do modelo vetorial, apresentada anteriormente.

4. Avaliação experimental

Visando uma avaliação preliminar 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 14 indivíduos para realizar o experimento. Simultaneamente, a base de dados contendo os artigos a serem recomendados foi carregada através do *harvesting* de metadados de todos os artigos cadastrados na BDBComp até junho de 2006, totalizando 3978 artigos de 113 edições de conferências. Com base nas informações dos últimos três anos (incluindo o ano corrente, ou seja, 2003 a 2006) presentes em tais currículos, foram gerados pelo sistema de recomendação 20 artigos para cada indivíduo. Além disso, foram apresentadas algumas informações, presentes na base, sobre cada um dos 20 artigos recomendados, extraídas diretamente dos metadados disponibilizados na BDBComp, tais como: título do artigo (*dc:title*), autores (*dc:creator*), link para o artigo completo (*dc:identifier*), idioma (*dc:language*), ano de publicação (*dc:date*), evento em que o mesmo foi publicado (*dc:source*) e resumo/abstract (*dc:description*).

As recomendações personalizadas foram geradas em ordem decrescente de grau de relevância relativo calculado pelo sistema (o artigo com 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) e foram enviadas aos respectivos indivíduos, sendo solicitado que cada um avaliasse as recomendações recebidas. Para tanto, cada indivíduo deveria atribuir um dos cinco conceitos (*Péssimo*, *Ruim*, *Médio*, *Bom* ou *Ótimo*), para avaliar a recomendação de cada um dos 20 artigos. Além disso, foi solicitado que os mesmos fizessem comentários sobre as recomendações. Os resultados obtidos, com base nas avaliações, serão apresentados e analisados na seção seguinte.

5. Resultados obtidos e conclusões preliminares

A Figura 3 apresenta um panorama geral das 14 avaliações recebidas, considerando duas situações distintas. Na primeira delas (Figura 3 (a)) é apresentada a porcentagem de cada uma das qualificações (na Figura representadas como *Péssimo*, *Ruim*, *Médio*, *Bom*, *Ótimo* e *Próprio Autor*), considerando o número de artigos recomendados (Top 20, Top 10, Top 5 e Primeiro). Analisando estes resultados, observamos que, considerando apenas o primeiro artigo recomendado para cada usuário, a porcentagem da qualificação *Ótimo* é a mais alta (42,86%). Além disso, muitos artigos dos próprios autores foram recomendados (28,57%). Estes resultados sugerem que o sistema está identificando adequada-

mente o perfil do usuário, especialmente pelo percentual de recomendações aos próprios autores. Nada melhor que um artigo do próprio autor para representar seus interesses de pesquisa. Também observamos que nenhuma qualificação *Péssimo* foi obtida para o primeiro artigo recomendado. Este resultado reforça a capacidade do sistema em fazer recomendação de forma ajustada aos interesses do usuário.

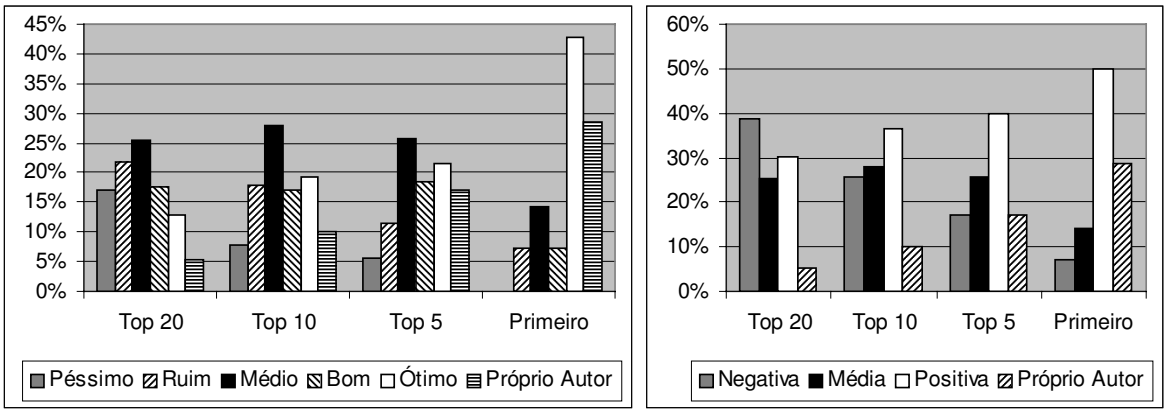

Figura 3. Resultados obtidos pela avaliação experimental (a) Gráfico Completo (b) Gráfico Resumido

Na situação mostrada na Figura 3 (b), foi reduzido o número de categorias na tentativa de se obter uma melhor visualização dos resultados. Aqui os conceitos *Péssimo* e *Ruim* foram agrupados em uma única categoria, referenciada como *Negativa* e os conceitos *Ótimo* e *Bom* foram agrupados em uma única categoria, referenciada como *Positiva*. Assim, podemos observar que as qualificações *positivas* obtidas são superiores às qualificações *negativas* na maioria dos casos, chegando a 50% de qualificações *positivas* contra apenas 7,14% de *negativas* se for considerado apenas o primeiro artigo recomendado. As qualificações *positivas* foram inferiores apenas quando consideramos os 20 artigos recomendados. Este resultado igualmente sugere um bom desempenho do sistema de recomendação.

O fato de que a BDBComp, atualmente, tem uma cobertura limitada em virtude de não cobrir todas as áreas da Ciência da Computação, influenciou na qualidade das recomendações. Esta observação foi realizada com base nos comentários feitos pelos próprios professores que participaram do experimento, tais como:

> *"... 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, as 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..."*

Além disso, no experimento realizado, autores que mudaram de área nos últimos três anos, podem ter qualificado negativamente publicações 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 gerar recomendações consistentes. Existem alguns indícios a 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."*

Por fim, na Figura 3 também pudemos observar que conforme o número de artigos considerados aumenta, diminui a qualificação dos mesmos. Isto já era esperado já que 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. Além disso, o número de artigos do próprio autor é menor à medida que mais artigos são recomendados, isso também era esperado, já que artigos do próprio autor devem ser os primeiros a ser recomendados por estarem mais próximos ao seu próprio perfil. Claro que para fins de testes, a análise da recomendação de artigos do próprio autor é bem interessante, mas na prática, tais artigos devem ser filtrados da recomendação.

Em uma outra análise, optamos por agrupar os usuários pesquisados em duas categorias: de professores e alunos da UFRGS. A Figura 4 apresenta as porcentagens de cada uma das qualificações recebidas (*Péssimo*, *Ruim*, *Médio*, *Bom*, *Ótimo* e *Próprio Autor*), por categoria de usuário, para as Top 10 e Top 5 recomendações. Com esta divisão pudemos observar que o sistema recebeu mais 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. Dessa forma, podemos supor também 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.

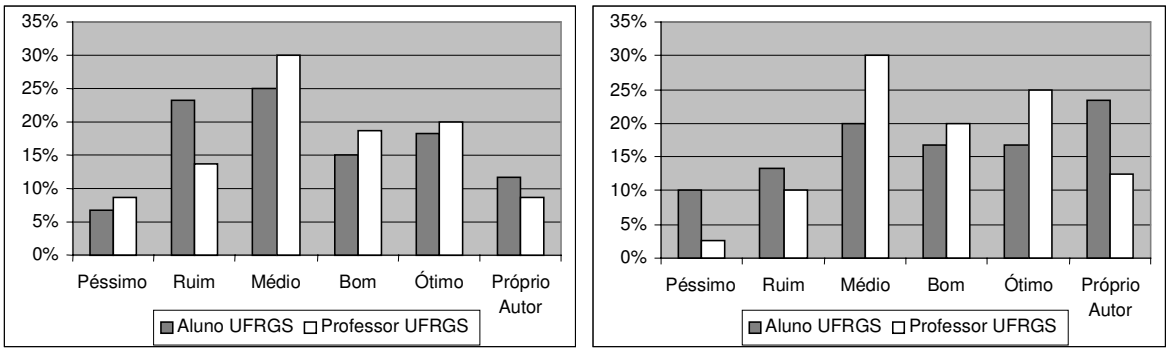

Figura 4. Resultados obtidos por categorias Aluno e Professor da UFRGS (a) Top 10 (b) Top 5

6. Considerações Finais e Trabalhos Futuros

O presente trabalho está inserido no contexto de diversos projetos de pesquisa, relacionados com bibliotecas digitais e com ensino a distância - EAD. A origem deste trabalho está associada com a recomendação de artigos para o suporte ao ensino. Nosso objetivo, ao final deste trabalho, é oferecer um *Web service* que permita sistemas de EAD oferecerem para os alunos recomendações dos artigos, disponíveis na Biblioteca Digital da SBC, mais adequados à complementação de aulas. Desta forma, os alunos disporão de material atualizado para complementarem seus estudos.

O estágio atual do sistema de recomendação proposto (protótipo) já permitiu uma experiência inicial de recomendação de artigos publicados nos Simpósios Brasileiros da SBC. O grande interesse da comunidade local por este tipo de serviço, que ficou demonstrado pela alta taxa de resposta à avaliação e pelos comentários dos participantes, nos estimulou a disponibilizar tão rápido quanto possível uma versão Web do serviço.

Referências Bibliográficas

Berners-Lee, Tim. (1999) "Weaving the Web". Harper, SF.

Callan, Jamie et al. (2003) "Personalisation and Recommender Systems in Digital Libraries". Joint NSF-EU DELOS Working Group Report. Maio.

CLEF (2005) "CLEF and Multilingual information retrieval", http://www.unine.ch/info/clef/, Institut interfacultaire d'informatique, University of Neuchatel.

DC-OAI (2005) "A XML schema for validating Unqualified Dublin Core metadata associated with the reserved oai_dc metadataPrefix", http://www.openarchives.org/OAI/2.0/oai_dc.xsd, Março.

Dublin Core (2005) "Dublin Core Metadata Initiative", http://dublincore.org, Setembro.

Ferneda, Edberto. (2003) "Recuperação de Informação: Análise sobre a contribuição da ciência da computação para a ciência da informação". Tese de Ciências da Comunicação. USP, São Paulo, cap. 4, p. 20-53.

Gutteridge, C. (2002) "GNU EPrints 2 overview", Jan. 01.

Huang, Z. et. al. (2002) "A Graph-based Recommender System for Digital Library". In: JCDL'02. Portland, Oregon.

Laender, A. H. F.; Gonçalves, M. A.; Roberto, P. A. (2004) "BDBComp: Building a Digital Library for the Brazilian Computer Science Community". In: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, Tuscon, AZ, USA, pp. 23-24.

Lattes-CNPq (2005) "Plataforma Lattes - Conselho Nacional de Desenvolvimento Científico e Tecnológico", http://lattes.cnpq.br/, Março.

LPML-CNPq (2005) "Padronização XML: Curriculum Vitae", http://lmpl.cnpq.br/lmpl/?go=cv.jsp, Março.

Maly, K.; Nelson, M.; Zubair, M.; Amrou, A. ; Kothamasa, S.; Wang, L.; Luce, R. (2004) "Light-weight communal digital libraries". In Proc. of JCDL'04, pages 237-238, Tucson, AZ.

OAI (2005) "Open Archives Initiative", http://openarchives.org, Outubro.

OAI-PMH (2005) "The Open Archives Initiative Protocol for Metadata Harvesting", http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm, Novembro.

Salton, Gerard; Buckley, Christopher. (1988) "Term-Weighting Approaches in Automatic Text Retrieval", Information Processing and Management: an International Journal, Volume 24, Issue 5, p. 513-523.

Sompel, H. V. de; Lagoze, C. (2000) "The Santa Fe Convention of the Open Archives Initiative". D-Lib Magazine, [S.l.], v.6, n.2, Feb.

Tansley, R.; Bass, M.; Stuve, D.; Branschofsky, M.; Chudnov, D.; McClellan, G.; Smith, M. (2003) "DSpace: An institutional digital repository system". In Proc. Of JCDL'03, pages 87-97, Houston, TX.