UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA

Marcus Fillipi Rosso Casagrande

TÉCNICA DE RECOMENDAÇÃO PARA REPOSITÓRIOS DIGITAIS BASEADA EM METADADOS E AGRUPAMENTO DE USUÁRIOS

Florianópolis

2014

Marcus Fillipi Rosso Casagrande

TÉCNICA DE RECOMENDAÇÃO PARA REPOSITÓRIOS DIGITAIS BASEADA EM METADADOS E AGRUPAMENTO DE USUÁRIOS

Dissertação submetida ao Programa de Pós Graduação em Ciências da Computação da Universidade Federal de Santa Catarina para a obtenção do Grau de Mestre em Ciências da Computação.
Orientador: Prof. Dr. Roberto Willrich.

Florianópolis

2014

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Casagrande, Marcus Fillipi Rosso
Técnica de recomendação para repositórios digitais baseada em metadados e agrupamento de usuários / Marcus Fillipi Rosso Casagrande ; orientador, Roberto Willrich - Florianópolis, SC, 2014.
77 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Ciência da Computação.

Inclui referências

1. Ciência da Computação. 2. Ciências da Computação. 3. Sistemas de Recomendação. 4. Filtragem Colaborativa. I. Willrich, Roberto. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. III. Título.

Marcus Fillipi Rosso Casagrande

TÉCNICA DE RECOMENDAÇÃO PARA REPOSITÓRIOS DIGITAIS BASEADA EM METADADOS E AGRUPAMENTO DE USUÁRIOS

Esta Dissertação foi julgada adequada para obtenção do Título de "Mestre em Ciências da Computação", e aprovada em sua forma final pelo Programa de Pós Graduação em Ciências da Computação da Universidade Federal de Santa Catarina.

Florianópolis, 28 de Fevereiro de 2014.

_____
Prof. Ronaldo dos Santos, Dr.
Coordenador do Curso

Banca Examinadora:

_____
Prof. Roberto Willrich, Dr.
Orientador, UFSC

_____
Prof. Ronaldo dos Santos Melo, Dr.
PPGCC - UFSC

_____
Prof. Luiz Otávio Alvarez, Dr.
PPGCC - UFSC

_____
Prof. José Valdeni de Lima, Dr.
PPGIE – UFRGS

Este trabalho é dedicado aos meus pais pelo apoio ao meu crescimento profissional.

AGRADECIMENTOS

Ao meu orientador Dr. Roberto Willrich, pela constante ajuda durante todo o andamento do mestrado; à Katiana Castro, por sempre me deixar a par das minhas atividades no PPGCC; ao meu grande amigo André Albino Pereira, pelo apoio e incentivo a este caminho que segui; e ao Projeto PRONEX Autores e Obras Catarinenses em Meio Digital, pelo grande esforço no desenvolvimento da Biblioteca Digital de Literatura.

RESUMO

Repositórios Digitais (RD) fornecem meios para armazenar, gerenciar e disseminar conteúdos virtuais. Estes sistemas adotam técnicas visando um melhor gerenciamento de seu conteúdo. Em particular, o uso de metadados auxilia na classificação e identificação dos conteúdos disponíveis, facilitando sua localização. Com a proliferação da informação digital, os RDs têm sido cada vez mais disseminados na Web. Mesmo com os conteúdos digitais organizados, o aumento constante nas coleções dos RDs tem causado preocupação quanto à facilidade do usuário em localizar conteúdos relevantes e ao mesmo tempo mantendo uma complexidade e tempo de processamento em um nível aceitável. Este trabalho propõe uma técnica de recomendação simples e escalável aplicada a repositórios digitais. Tal técnica se baseia na construção do perfil do usuário de maneira implícita, observando os valores de um certo conjunto de metadados. A recomendação proposta faz uso de uma forma de agrupamento, onde usuários são agrupados de acordo com os valores de metadados mais abundantes em seus perfis, reduzindo o tempo de processamento da Filtragem Colaborativa aplicada no sistema. A recomendação proposta foi testada e aplicada numa biblioteca digital de obras literárias.

Palavras-chave: Sistema de Recomendação, Perfil de Usuário, Clusterização, Agrupamento, Repositório Digital, Metadado.

ABSTRACT

Digital repositories (DR) provide ways to store, manage and disseminate virtual content. These systems make use of techniques aimed at better management of their content. In particular, the use of metadata assists in the classification and identification of available content, facilitating their location. With the proliferation of digital information, DRs have been increasingly widespread on the Web. Even with organized digital content, the constant increase in collections of DRs has caused concerns about the ease of the user locating relevant content and at the same time maintaining the complexity and the processing time in an acceptable level. This work proposes a recommendation technique that is simple and scalable, applied to digital repositories. This technique relies on the construction of the user profile in an implicit manner, observing the values of a certain set of metadata. This Recommendation makes use of a form of grouping in which users are grouped according to the most abundant metadata values in their profiles, reducing the processing time of the Collaborative Filtering applied in the system. The proposed recommendation has been tested and applied in a digital library of literary works.

Keywords: Digital Repository, Recommender Systems, User Profile, Clustering, User-subgroups.

LISTA DE FIGURAS

LISTA DE ABREVIATURAS E SIGLAS

RD .....Repositório Digital
DC .....Dublin Core
BD-LB.....Biblioteca Digital de Literatura Brasileira
FC.....Filtragem Colaborativa
FBC .....Filtragem Baseada em Conteúdo
RD .....Repositório Digital
OA.....Objetos de Aprendizagem
PE .....Preferências Específicas
CAR.....Controle de Apresentação dos Resultados
RIA.....Recuperação de Informações Adaptativas
TVI.....TV Interativa
MO .....Metadados Observados

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Repositórios de conteúdo digitais são sistemas, geralmente Web, que oferecem uma série de serviços, principalmente para o depósito, organização e acesso a conteúdos digitais, construídos de diferentes formas e com diferentes propósitos. Repositórios Digitais geralmente abrigam informações relacionadas a um escopo. Por exemplo, repositórios que abrigam artigos científicos, livros educativos, ou mesmo materiais didáticos.

Existem diversos termos utilizados para referenciar um repositório de conteúdos digitais, como Repositório Digital (RD), Biblioteca Digital ou Biblioteca Virtual. Este trabalho considera que Bibliotecas Digitais ou Virtuais são repositórios construídos sobre princípios rígidos de gerenciamento de informação aplicados em biblioteconomia por centenas de anos. O termo RD é adotado neste trabalho como um termo mais genérico, que é usado para referenciar sistema que, além das Bibliotecas Digitais, englobaria repositórios que não se baseiam em técnicas rígidas de gerenciamento de informação.

A organização do conteúdo em RDs é baseada na definição de valores a um conjunto de descritores destes conteúdos, chamados de metadados. Em geral, metadados são definidos como dados sobre os dados. Existem algumas iniciativas de padrões de metadados para repositórios, sendo que um dos principais são os metadados Dublin Core (DC) (Dublin Core, 2013) e IEEE LOM (IEEE, 2004).

O conteúdo disponibilizado por um RD pode representar a informação na forma de diferentes mídias, como texto, imagens e vídeos. Em um contexto educacional, por exemplo, as coleções oferecidas pelos RDs são compostas por conteúdos, ou objetos, de aprendizagem.

É crescente o número de RDs na forma de repositórios institucionais e bibliotecas digitais. Cresce também o tamanho das coleções disponibilizadas por estes repositórios. Apesar dos repositórios oferecerem meios para a organização dos conteúdos disponibilizados, com o aumento do número de conteúdos, o usuário se depara novamente com o problema clássico de sobrecarga de informação, quando há um grande esforço por parte do usuário para localizar a informação desejada dentro de um conjunto enorme de informações disponíveis.

Existem diversas técnicas para aumentar a eficiência na recuperação da informação. Algumas técnicas de recuperação personalizada de informações (Maleki-Dizaji et al., 2003), (Willrich et

al., 2006) ajudam no aprimoramento do acesso à informação. De maneira geral, estas técnicas permitem tratar uma enorme quantia de dados e filtrá-los de acordo com as preferências e interesses do usuário e da comunidade de usuários.

Um exemplo de RD oferecendo sistemas de recuperação de informações personalizadas é a Biblioteca Digital de Literatura Brasileira (BD-LB) (www.literaturabrasileira.ufsc.br). O usuário da BD-LB se cadastra no sistema e seu perfil é construído automaticamente via análise dos valores de um conjunto de metadados dos conteúdos acessados por ele. Baseado no perfil, os resultados das buscas são reordenados levando em conta a relevância dos conteúdos quanto à satisfação do critério de busca e também a preferência dos documentos calculada com base no perfil do usuário e valores de metadados do conteúdo (Furtado et al., 2009).

Outra forma usual para melhoria do acesso à informação é o uso de técnicas de recomendação (Basu et al., 1998), (Breese et al.,1998) e (Cazella et al., 2008). Sistemas de recomendação usam informações de utilização do RD ou preferências do usuário para identificação de conteúdos que sejam possivelmente de interesse do usuário que faz a busca no sistema. Estes tipos de sistemas são muito utilizados em sites de comércio eletrônico. Por exemplo, Amazon (www.amazon.com) e eBay (www.ebay.com), utilizam técnicas de recomendação para sugerir produtos relacionados ao conteúdo que o usuário está adquirindo, visando o aumento das vendas.

Sistema de recomendação é um tipo particular de filtragem de informação, cujo objetivo é apresentar ao usuário um conjunto de itens (p.e., produtos ou conteúdos digitais de qualquer natureza) que sejam susceptíveis de serem de interesse do usuário. Várias técnicas são empregadas, sendo as mais tradicionais a Filtragem Colaborativa (FC) (Herlocker et al., 2004), (Herlocker et al., 2012), (Huang et al., 2005) e a Filtragem Baseada em Conteúdo (FBC) (Mooney; Roy, 1999), (Martínez et al., 2007). Também existem técnicas de Filtragem Híbridas, que geralmente fazem uso das técnicas de FC e FCB juntas a fim de buscar as vantagens de ambas.

As técnicas de FBC geralmente analisam os atributos dos itens que se demonstraram interessantes para um determinado usuário para recomendar a ele outros itens similares. Como apresentado mais adiante, uma das desvantagens da FCB se dá pelo fechamento dos itens a serem recomendados.

Em geral, as técnicas de FC buscam recomendar conteúdos acessados por usuários com perfis semelhantes ao usuário foco da

recomendação. Como visto mais adiante, essas técnicas possuem dois problemas bem conhecidos: o problema de esparsidade (quando não há recomendações para itens impopulares devido à baixa quantidade de acessos a tais itens por outros usuários), e de escalabilidade (Sarwar et al., 2002), (Claypool et al., 1999), (Park et al., 2012).

Uma das formas para tratar o problema da escalabilidade é a redução do espaço de busca pelo agrupamento de usuários (Sarwar et al., 2002), (Kim; Yang, 2005) ou agrupamento de itens (O'Connor; Herlocker, 2001), (Li; Kim, 2003). Estas técnicas buscam reduzir a complexidade da recomendação sem interferir em sua qualidade.

Com o avanço das tecnologias de informações, novos paradigmas vêm sendo aplicados aos sistemas de recomendação, como o uso de ontologias e web semântica, que ajudam na modelagem e interpretação dos dados utilizados no cálculo da recomendação.

1.1 PERGUNTA DE PESQUISA

Este trabalho busca responder a pergunta a seguir: **Como desenvolver uma técnica eficaz de recomendação voltada a RDs?**

Embora neste trabalho o cenário de aplicação e testes realizados seja no contexto de RDs, a solução proposta busca a utilização de diferentes técnicas voltadas à sistemas que fazem uso de metadados para descrever seus conteúdos. Tais técnicas se destacam por sua simplicidade, exigindo baixo custo computacional sem comprometer a qualidade do conteúdo buscado para recomendação.

1.2 JUSTIFICATIVA

Várias técnicas de recomendação vêm sendo utilizadas e aprimoradas, tendo grande utilização nos RDs e principalmente nos sistemas de comércio eletrônico.

RDs têm abrigado cada vez mais informações de diversos tipos. Com o aumento do conjunto de conteúdos digitais disponíveis, faz-se necessário o uso de ferramentas que facilitem a busca por conteúdos relevantes, que são aqueles que atendem as demandas ou interesse de um usuário em particular.

Para auxiliar na modelagem do perfil dos usuários, muitos RDs fazem uso de uma construção explícita dos perfis. Esta forma de construção do perfil é a forma mais adequada, pois o usuário realimenta o sistema indicando explicitamente seus interesses. Mas infelizmente,

esta técnica possui uma limitação severa, pois os usuários devem realizar um esforço extra para informar suas preferências pessoais. Na maior parte das técnicas de recomendação, o usuário deve realizar uma avaliação dos conteúdos que acessa. Devido a esta carga extra de trabalho, muitos usuários tendem a não responder a tal construção do perfil, como constatado por (WeiWei et al., 2009) e (Speretta; Gauch, 2005).

A grande maioria dos sistemas de recomendação não explora os valores dos metadados de conteúdos para estimar o perfil do usuário. Utilizando de uma abordagem diferente da maioria dos sistemas propostos, a técnica aqui proposta adota uma abordagem de construção implícita do perfil do usuário através do uso dos valores de metadados dos conteúdos acessados por ele ao longo do tempo.

Fazendo uso de tal modelagem de perfil, foi proposto um sistema que auxilia na busca por conteúdos relevantes em RDs. Através da adoção de uma técnica de recomendação com agrupamento de usuários, pretende-se reduzir o tempo de processamento exigido na busca por tais conteúdos sem reduzir a qualidade da informação trazida.

1.3 OBJETIVOS

Em sequência à pergunta de pesquisa e motivação, são apresentados os objetivos geral e específicos deste trabalho.

1.3.1 Objetivo Geral

Este trabalho tem por objetivo descrever, implementar e testar uma técnica de recomendação de conteúdos em RDs visando reduzir o tempo despendido pelo sistema sem prejudicar na qualidade da recomendação. Tal técnica é voltada para um RD, pois ela faz uso de metadados de conteúdos textuais para modelar o perfil dos usuários. Isso é feito através da observação de seus históricos de acesso e os valores de metadados dos conteúdos.

Para tratar o problema de escalabilidade a técnica de recomendação proposta adota um mecanismo de agrupamento de perfis de usuários com interesses similares. A partir da identificação dos grupos de interesse na qual o usuário foco da recomendação pertence, é gerada a lista de recomendação com base nos vizinhos próximos do usuário em todos os grupos a que o usuário foco da recomendação pertence.

Uma das principais vantagens da técnica proposta é a sua simplicidade computacional, devido à análise dos valores dos metadados do perfil do usuário para a separação dos usuários em diferentes grupos de interesse. Além disso, outra vantagem é a redução do problema de escalabilidade da recomendação, via a técnica de agrupamento definida.

A técnica proposta foi implementada na BD-LB, com a qual foram realizados testes de qualidade e escalabilidade da recomendação.

1.3.2 Objetivos Específicos

A partir do objetivo geral, pôde-se formular os seguintes objetivos específicos:

1. Estudar as características e ferramentas mais relevantes que estruturam um RD.
2. Analisar as principais técnicas de recomendação para Repositórios Digitais.
3. Definir e desenvolver uma técnica de recomendação que se beneficie do uso de metadados existentes num RD.
4. Definir um modelo de perfil de usuário com base nos valores dos metadados das obras acessadas.
5. Definir uma técnica de recomendação visando reduzir a escalabilidade do processo de busca de conteúdos relevantes ao usuário. Sendo que a proposta deverá utilizar técnica de agrupamento visando aumentar a sua escalabilidade.
6. Avaliar a proposta através de técnicas que simulam grandes quantidades de usuários para verificar o tempo de processamento utilizado requerido, e utilização de usuários reais para verificar a qualidade da recomendação proposta.

1.4 ESTRUTURA DO TRABALHO

Esta dissertação está organizada da forma que segue. O capítulo 2 introduz os principais conceitos relacionados a RDs. O capítulo 3 apresenta um levantamento do estado da arte em Sistemas de Recomendação, descrevendo alguns dos principais tipos de sistemas de recomendação existentes. O capítulo 4 mostra como foi modelado o conceito proposto neste artigo. O capítulo 5 relata as pesquisas e testes comprovando a eficiência da técnica adotada. Finalmente, o capítulo 6 apresenta as conclusões e propostas para trabalhos futuros.

2 REPOSITÓRIOS DIGITAIS

Com a vasta quantidade de informações que se tem hoje em dia graças à globalização e a facilidade de expansão obtida através da Internet e da Web, a troca e a disponibilização de informações se tornou mais prática, rápida e eficiente. Vários setores da sociedade têm utilizado as tecnologias digitais para criar ambientes virtuais para disseminar e tornar disponível conteúdos digitais através da Web.

Conteúdos digitais podem ser disseminados de várias formas. Uma delas é através da criação de RDs, que armazenam coleções de conteúdos digitais, geralmente voltados a um escopo específico. Este tipo de sistema tem tido amplo crescimento ao longo dos últimos anos. Apenas a OpenDOAR (www.opendoar.org), um diretório de repositórios acadêmicos de acesso livre, conta com 2522 repositórios espalhados pelo mundo cadastrados em seu domínio.

Este capítulo apresenta os principais conceitos relacionados aos RDs. Seguida a definição de RDs, este capítulo descreve o conceito de metadados, que são amplamente utilizados nos RDs. Na sequência, ele apresenta um dos problemas que têm sido cada vez mais frequentes nos RDs, o problema da sobrecarga de informações. Finalmente, este capítulo apresenta o RD onde a técnica proposta nesta dissertação foi implementada e testada.

2.1 DEFINIÇÃO DE REPOSITÓRIO DIGITAL

Vários autores definem Repositórios Digitais de diferentes formas. Por exemplo, em (Heery; Anderson, 2005) os autores definem o termo “repositório” como sendo uma coleção de objetos digitais, especializando o termo Repositório Digital para aqueles repositórios de coleções digitais contendo as seguintes características:

- O depósito do conteúdo no RD pode ser realizado pelo criador, proprietário ou qualquer outro usuário.
- O repositório gerencia os conteúdos e seus metadados.
- O RD deve oferecer um conjunto de serviços para seus usuários, como mecanismos de busca, depósito, recuperação e controle de acesso dos conteúdos disponibilizados.
- O RD precisa ser confiável e bem gerenciado: suas informações devem ter veracidade de acordo com suas fontes e seu conteúdo bem organizado.

Não há uma clara diferenciação entre vários conceitos similares ou sinônimos de Repositórios Digitais. Em (Cleveland, 1998), os autores afirmam que há muita confusão em torno dessas definições. Segundo a percepção deste autor, a comunidade de biblioteconomia tem usado várias definições distintas ao longo dos anos para denotar tais conceitos, incluindo "biblioteca eletrônica", "biblioteca virtual", e até "biblioteca sem paredes"; e que nunca esteve exatamente claro suas diferenças de significado. Em (Cleveland, 1998), o autor afirma que o termo Biblioteca Digital é o termo mais aceito e difundido.

Diferentes autores usam diferentes definições e geralmente elas se baseiam em opiniões pessoais (que não são estabelecidas por instituições normalizadoras ou organizações) sobre os termos utilizados. Em geral, tais termos são considerados muito semelhantes. (Duncan, 2003), apesar de afirmar que a utilização do termo biblioteca se aplica bem ao repositório digital, define que uma biblioteca é onde os recursos estão armazenados e apenas os bibliotecários mantêm o controle do que pode ser armazenado ou incluído na biblioteca. Já os repositórios enfatizam mais o fato de que as pessoas podem contribuir com recursos adicionais ou objetos de aprendizado que são distribuídos na comunidade.

Algumas federações de bibliotecas digitais e iniciativas na área têm suas próprias definições de bibliotecas digitais. A DLF (*Digital Library Federation*) (http://www.diglib.org) define bibliotecas digitais como uma organização que oferece os recursos comuns de bibliotecas, tais como material de consulta, documentos, além de uma equipe própria e especializada, responsável por organizar, coordenar e dar assistência aos documentos digitais existentes.

Por sua vez, a DLI (*Digital Libraries Initiative*) (http://www.dlib.org/) afirma que uma Biblioteca Digital deve fornecer serviços para a continuidade, melhoria e difusão de seu conjunto de informação, além de seu conjunto de dados e organização.

Outro termo muito usado atualmente é o de "Repositório Institucional". Trata-se de um RD que busca preservar e disseminar o conhecimento (em formato digital) de uma instituição, ou comunidade universitária (Crow, 2002).

A presente dissertação adota o termo Repositório Digital (RD), pois o autor considera este termo mais genérico que Biblioteca Digital. Para este trabalho o termo RD engloba repositórios institucionais e também biblioteca digital, que são construídos seguindo os princípios rígidos de gerenciamento de informação aplicados em biblioteconomia por centenas de anos.

2.2 METADADOS

RDs adotam metadados para caracterização das coleções disponibilizadas. Metadados significam dados sobre dados. São dados que fornecem informações sobre um ou mais aspectos de dados como criador, data de criação, assunto, tipo, etc.

O uso de metadados tem por objetivo permitir a catalogação do conteúdo no RD, facilitando assim a busca e organização destes conteúdos. Segundo (Furtado et al., 2009), no processo de recuperação da informação é possível aos usuários dos RDs realizarem consultas booleanas para especificar os valores de elementos de metadados desejados dos recursos que o usuário deseja recuperar.

Para melhorar a interoperabilidade entre os RDs, vários padrões de metadados vem sendo adotados e aprimorados. Os principais padrões na área são o Dublic Core e IEEE LOM.

2.2.1 Iniciativa de Metadados Dublin Core

A Iniciativa de Metadados Dublin Core DCMI (Dublin Core Metadata Iniciative, 2014) é uma organização aberta que estimula a inovação, padronização e práticas de usos de metadados descritores de recursos digitais. Suas atividades incluem discussões sobre a modelagem e arquitetura, conferências globais e esforços educacionais para promover a aceitação dos padrões de metadados e seus melhores usos. A DCMI definiu o conjunto DCES (*Dublin Core Metadata Element*), um conjunto mínimo de 15 elementos capazes de descrever recursos digitais. Estes 15 elementos, sendo todos recomendados, mas nenhum obrigatório, são:

- *Audience*: A quem o conteúdo se destina
- *Contributor*: Uma entidade que tenha contribuído para o conteúdo.
- *Coverage*: Uma área espacial ou temporal de abrangência do conteúdo.
- *Creator*: Uma entidade responsável pela criação do conteúdo.
- *Date*: Período de tempo associado com a disponibilização do conteúdo.
- *Description*: Descrição do conteúdo (*abstract*, tabela de conteúdos...).
- *Format*: Formato do conteúdo, ou dimensões como tamanho e duração.

- *Identifier*: Uma referência única que identifique o conteúdo num dado contexto.
- *Language*: A linguagem do conteúdo.
- *Publisher*: Entidade responsável por tornar o conteúdo disponível.
- *Relation*: Um recurso com o qual o conteúdo esteja relacionado.
- *Rights*: Informação quanto a propriedade intelectual ou direito autorais do conteúdo.
- *Source*: Recursos relacionados dos quais se deriva o conteúdo em questão.
- *Subject*: Assunto do qual o conteúdo trata.
- *Title*: O nome dado ao conteúdo.
- *Type*: O tipo, natureza, gênero, do conteúdo.

2.2.2 Metadados IEEE LOM

O formato IEEE LOM (*Learning Object Metadata*) (IEEE, 2003) é mais voltado ao contexto educacional, sendo um padrão que especifica a sintaxe e semântica dos metadados visando descrever Objetos de Aprendizagem (OA). O IEEE define um OA como qualquer entidade, digital ou não, que possa ser usado para aprendizado, educação ou treinamento.

Segundo (IEEE, 2003), os propósitos do IEEE LOM incluem características como:

- Possibilitar a procura, aquisição e utilização dos OA.
- Possibilitar a distribuição e troca dos OA entre tecnologias suportadas pelos sistemas de aprendizado.
- Possibilitar que agentes computacionais possam compor conteúdos personalizados para um usuário individual.
- Suportar uma economia crescente para OA que suporte todas as formas de distribuição voltadas ou não ao lucro.
- Possibilitar que quaisquer instituições possam expressar conteúdo educacional em formato padronizado que não dependa do conteúdo descrito em si.
- Definir um padrão que seja simples, porém estendível a múltiplos domínios e jurisdições facilitando sua adoção.
- Suporte a segurança e autenticação para distribuição e uso de OA.

2.3 SOBRECARGA DE INFORMAÇÃO

Os metadados permitem a organização das coleções disponíveis nos RDs, facilitando a busca e navegação no conteúdo. O uso de metadados permite agregar, identificar e localizar conteúdos relevantes ao usuário que busca por conteúdos. Apesar disto, muitas buscas em RDs podem gerar imensas quantidades de resultados caso o montante de recursos digitais disponíveis seja grande e/ou o critério de busca adotado pelo usuário seja pouco específico.

De acordo com (Wan; Liu, 2008), há três métodos de se buscar informações em um RD. O primeiro método é a navegação livre, onde o usuário navega por uma coleção para localizar uma informação específica. O segundo método, o mais popular, é a busca textual. Neste método, conteúdos baseados em textos são indexados para que o usuário possa fazer uso de termos ou palavras-chave. Tal método inclui as buscas por valores de metadados contidos no RD. O terceiro método é a busca baseada em conteúdo, que permite ao usuário filtrar conteúdos de imagens, áudio, ou vídeo. Algumas dessas características incluem cor, textura, tamanho ou movimentos. Um exemplo disso seria o uso de imagens numa busca de informações: um usuário busca certo tipo de material, como uma peça de mobília, porém não sabe o nome de tal material. Ele então utiliza uma imagem de tal material como critério de busca e o sistema reconhece imagens semelhantes junto com outras informações sobre tal material.

Independentemente do método, buscas com critérios pouco específicos realizadas nos RDs podem acabar resultando em um número muito grande de conteúdos que atendam ao critério de busca, principalmente em RDs com grandes coleções disponíveis. Tal cenário resulta no problema da sobrecarga de informações. A sobrecarga, no contexto deste trabalho, ocorre quando há uma quantidade de informações muito grande disponível numa busca, onde a quantidade de itens atendendo ao critério de busca é enorme, dificultando ao usuário encontrar o que estava procurando.

Um dos objetivos deste trabalho é a resolução deste problema. Através do uso da recomendação e de técnicas que reduzam o espaço de busca do repositório, busca-se identificar a quantidade mínima de resultados ordenados pela relevância que têm com o perfil do usuário que realiza a busca no sistema, reduzindo o problema da sobrecarga de informações.

2.4 BD-LB

A Biblioteca Digital de Literatura Brasileira (BD-LB) (www.literaturabrasileira.ufsc.br) conta atualmente com uma vasta coleção de obras literárias de domínio público e informações sobre escritores brasileiros. Estão catalogadas, no momento da escrita desta dissertação, 74.156 obras literárias (3.416 delas têm os textos integrais disponibilizados) de 18.238 autores diferentes.

A BD-LB é caracterizada pelo uso de código aberto, adotando a plataforma LAMP (Linux, Apache, MySQL, PHP). Existem diversos tipos de usuários: administradores, com a responsabilidade de configuração e personalização do sistema, bem como a atribuição de papeis aos usuários; colaboradores, com a função de cadastramento de dados sobre autores e obras literárias; e os usuários, que são alunos e professores de literatura, bem como leitores de obras literárias.

2.4.1 Metadados Descritores

Para a descrição dos conteúdos da BD-LB, além do título do conteúdo e autor, também há um conjunto de metadados descritores considerados relevantes para sua localização. Os conteúdos são descritos por:

- Título: Descreve o título do conteúdo acessado.
- Autores: Descreve um ou mais autores do conteúdo.
- Pseudônimo: Apelido pelo qual o autor é mais conhecido.
- Tipo: Descreve o tipo do conteúdo, que pode se estender além de obra literária, como nota de jornal, publicação na imprensa, etc.
- Gênero: o gênero literário do conteúdo descrito.
- Ano: descreve o ano ou século em que o documento foi produzido.
- Localização: A qual entidade cultural pertence determinado conteúdo.
- Descrição: Descrição breve a respeito do conteúdo.
- Data de inclusão: Data em que tal conteúdo foi inserido na BD-LB.
- Data de atualização: Última atualização do conteúdo.
- Tipo de documento: O tipo ou extensão do conteúdo.
- Idioma: O idioma em que o conteúdo está disponível.
- Editora: Nome da editora do conteúdo disponível.

2.4.2 Serviços de Busca e Navegação no Acervo

A BD-LB oferece uma série de interfaces de busca e navegação:

- Navegação no conteúdo: a BD-LB permite ao usuário navegar nos documentos, autores e acervos cadastrados na biblioteca. A navegação é realizada via o clique na letra inicial do título da obra ou nome do autor. No caso da navegação por acervo, o usuário simplesmente seleciona o acervo e suas obras são apresentadas em ordem alfabética.
- Busca Simples: onde o usuário fornece palavras-chave referentes aos autores ou obras, podendo indicá-las como frase exata ou palavras separadas. Esta busca é restrita a alguns metadados, como título, subtítulo, autor e pseudônimo.
- Busca Documento: fornece informações mais detalhadas a respeito da busca além do título, como a busca por metadados como: autores, gênero literário, período, dentre outros.
- Busca Autor: Permite a busca específica por autores cadastrados, onde se pode incluir informações como nome, pseudônimo, data de nascimento e falecimento. Tal busca é realizada sobre os valores de metadados descritivos dos autores das obras.
- Busca Conteúdo: Busca textual semelhante à busca simples, porém a busca é sobre o conteúdo das obras em formato textual.

A BD-LB conta também com uma seção de navegação, onde é possível navegar entre e autores e documentos em ordem alfabética, onde é possível selecionar a inicial desejada.

2.4.3 Recuperação Personalizada de Informações

Uma característica importante da BD-LB diz respeito ao cadastro e identificação dos usuários. Através do uso de valores de metadados que descrevem as obras disponíveis, a BD-LB possibilita a construção do perfil de acessos e preferências do usuário. Através disso, é possível fazer uso de técnicas de recuperação de informações que possibilitam uma busca mais focada nos interesses do usuário.

Usuários identificados pelo sistema têm seu perfil modelado com os registros de acessos a conteúdos através das frequências dos valores de metadados autor e gênero. Fazendo uso da Recuperação de Informações Adaptativas (RIA), o usuário identificado que faz uma busca tem seus resultados trazidos por ordem de relevância com seu perfil, que são mostrados por uma coluna extra na apresentação dos resultados de busca, identificada por 'Escore'. Tal coluna descreve, em forma de porcentagem dentre todos os resultados trazidos, qual o grau de relevância que tal conteúdo tem com o perfil do usuário que fez a busca. O perfil P_k do usuário u_k adotado pela BD-LB é definido por (Furtado et al., 2009):

$$P_k = (DPU_k, PG_k, PE_k) \quad (1)$$

Onde:

- Dados Pessoais do Usuário (DPU): são dados como dados para autenticação do usuário, seu nome completo, informações de contato e sua página Web. Tais dados são informados de forma explícita.
- Preferências Gerais (PG): são as preferências em termos de formato de apresentação e de conhecimentos gerais do usuário (por exemplo, a lista de línguas conhecidas). As PG devem ser explicitamente informadas pelo usuário.
- Preferências Específicas (PE): são as preferências do usuário relacionadas ao domínio da BD. Podem ser iniciadas de forma opcional pelo usuário e são atualizadas automaticamente de forma implícita baseando-se nos valores dos metadados dos conteúdos acessados pelo usuário.

Para a determinação dos pesos de relevância dos valores de metadados, a BD-LB adota a análise por frequência. Assim, o peso de relevância de determinado valor de metadado é dado pelo número de vezes que ele está presente nos conteúdos acessados pelo usuário.

A apresentação dos resultados das consultas na BD-LB mostra uma estrutura contendo vários campos de metadados, como descrito anteriormente. Com a recuperação personalizada de informações é possível que o usuário possa reordenar os resultados da busca de acordo com um dos termos que descrevem os resultados trazidos. P.ex., reordenar por título, autor(es), gênero, etc.

A BD-LB considera que cada conteúdo acessado contribui para a modelagem do perfil do usuário. Com isso, é assumido que o acesso a um conteúdo é um indicativo de interesse por tal conteúdo. Porém, um possível problema se dá pelo acesso precipitado, onde o usuário, após o acesso a determinado conteúdo, percebe que ele não é de seu interesse. Para reduzir este problema, a BD-LB não oferece acesso direto ao conteúdo após o clique, mas sim, uma descrição da obra inicialmente.

Mesmo com a possibilidade de visualização da obra antes de seu acesso, um usuário pode decidir que certo conteúdo ou valor de metadado não condiz com suas preferências. Por isso, a BD-LB oferece uma edição explícita do perfil, onde o usuário pode manualmente excluir certos valores de metadados que não considera como sendo preferenciais em suas buscas.

2.5 CONSIDERAÇÕES FINAIS

Neste capítulo foram vistos vários aspectos referentes aos RDs, como definições, uso de metadados e apresentação do RD onde o sistema proposto está implementado.

Este capítulo também abordou o problema da sobrecarga de informação, cada vez mais recorrente nos RDs. Esta dissertação busca tratar este problema através do uso da recomendação. O capítulo a seguir descreve vários aspectos referentes aos Sistemas de Recomendação, descrevendo várias técnicas estudadas existentes e exemplos de sua utilização em diferentes ambientes.

3 SISTEMAS DE RECOMENDAÇÃO

Sistemas de recomendação têm por objetivo apoiar o usuário na localização de conteúdos através da apresentação de uma lista de itens (geralmente conteúdos, produtos ou serviços) que sejam de interesse do usuário. Sistemas de recomendação são úteis quando o universo de escolhas é muito grande e/ou desconhecido, facilitando ao usuário a localização de recursos relevantes.

Vários autores têm suas definições para os sistemas de recomendação. Em (Resnick; Varian, 1997), uma das publicações mais referenciadas na área de Sistemas de recomendação, é definido que em um sistema típico de recomendação, se busca identificar em um conjunto de conteúdos (ou itens) aqueles que possam ser de interesse de um determinado usuário e em seguida apresentar a este usuário uma lista dos conteúdos identificados como de seu interesse. Conforme descrito por (Cazella et al., 2010), um dos grandes desafios dos sistemas de recomendação é realizar a combinação adequada entre as expectativas do usuário e os conteúdos ou serviços oferecidos a ele. O nível de relevância de cada conteúdo ou item para um dado usuário é determinado usando diferentes técnicas, dependendo da abordagem adotada.

Por sua vez, em (Ruotsalo, 2010) os autores definem que os sistemas de recomendação são um tipo específico de sistemas de filtragem de informação usados para identificar um conjunto de objetos que são relevantes para um usuário.

Existem diversas aplicações Web que incluem serviços de recomendação, como repositórios digitais e sistemas de comércio eletrônico. Além destes, alguns trabalhos de recomendação para programas de TV digital têm surgido (Santos; Ferraz, 2011), (Bernardo et al., 2011), (Bernhaupt et al., 2008), (Cremonesi; Turrin, 2010).

3.1 TÉCNICAS DE RECOMENDAÇÃO

Existem diversas técnicas de sistemas de recomendação, sendo que as principais são a Filtragem Colaborativa (Dan-Er, 2009), (O'Donovan et al., 2008), (Huang et al., 2005), Filtragem baseada em conteúdo (Boutemedjet; Ziou, 2008), (Mooney; Roy, 1999) e Filtragem Híbrida (Basilico; Hofmann, 2004), (Torres et al., 2004), (Vellino; Zeber, 2007). Existem outras técnicas menos populares não por seu nível de importância, mas em geral por seu recente desenvolvimento.

Esta seção apresenta em detalhes a Filtragem colaborativa, Filtragem baseada em conteúdo, bem como outras técnicas de sistemas de recomendação.

3.1.1 Filtragem Colaborativa

A FC é uma das técnicas mais amplamente utilizadas pelos sistemas de recomendação, inclusive aqueles de sites de comércio virtual (Sarwar et al. 2000), (Dan-Er, 2009).

Na FC (Dan-Er, 2009), (O'Donovan et al., 2008), (Huang et al., 2005), os conteúdos a serem recomendados a um usuário são determinados com base na identificação de grupos de usuários com perfis similares. Este grupo, chamado geralmente de "vizinhos próximos", é determinado para cada usuário através do uso de funções que determinam a similaridade entre avaliações ou acessos a conteúdos realizados pelo usuário foco da recomendação e pelos outros usuários do sistema. As técnicas de recomendação colaborativa assumem que existe uma grande probabilidade de que um usuário irá gostar de conteúdos que seus vizinhos próximos gostam.

Na FC o perfil do usuário mantém informações referentes ao usuário, como dados pessoais e dados que permitam identificar seus conhecimentos e interesses. Este perfil pode ser determinado de maneira explícita ou implícita:

- **Construção explícita**: o usuário explicitamente realiza uma avaliação de um dado documento após sua leitura, classificando assim o nível de relevância do documento avaliado. Apesar de se obter uma classificação mais direta e exata do usuário, essa maneira tem a desvantagem de necessitar que o usuário ceda parte do seu tempo para avaliar os itens disponíveis, fazendo com que poucos usuários se proponham a fazer tal avaliação. Em geral, não há uma recompensa direta para o usuário que fornece avaliações, de forma que isso ajudaria a outras pessoas (Middleton et al. 2004).
- Construção implícita: os documentos têm seu nível de relevância inferido de acordo com o comportamento do usuário. Estas técnicas utilizam o histórico de acesso e leitura do documento pelo usuário, da forma de navegação no sistema e os tipos de buscas realizadas para inferir o perfil do usuário. Essa forma tem a vantagem de ser feita automaticamente, sem a necessidade de prestação de

intervenção direta do usuário. Como aspecto negativo, ela tende a ser menos precisa que a construção explícita.

(Middleton et al. 2004) na sua definição de FC, ignora a possibilidade de construção implícita do perfil de usuário. O autor define a FC como aquela que realiza a recomendação através da requisição de que os usuários avaliem itens explicitamente, e então recomenda novos itens que usuários semelhantes tenham avaliado positivamente. Trabalhos mais recentes consideram também a construção implícita do perfil de usuário.

Algumas propostas realizam uma combinação das duas abordagens acima para formar o perfil do usuário. Estas propostas geralmente fazem a construção do perfil do usuário de maneira implícita inicialmente, possibilitando uma avaliação explícita para "remodelagem" do perfil após a recomendação. Dessa forma o usuário tem a capacidade de avaliar a recomendação recebida, assim, informando o sistema sobre seus gostos e atualizando seu perfil.

Algoritmos de FC podem ser divididos em duas categorias (Breese et al.,1998), (Boutemedjet; Ziou, 2008):

- **Algoritmos baseados em memória**: o cálculo da similaridade dos perfis de usuários é feito comparando-os às avaliações feitas por outros usuários similares, através de algoritmos de comparação que verificam o grau de similaridade entre perfis de usuários. Um exemplo típico são os mecanismos que estimam os vizinhos próximos do usuário alvo da recomendação. Essa técnica é mais simples e popular, seus dados são facilmente adicionados, mas pode ter seu desempenho comprometido com o crescimento da base de dados;
- **Algoritmos baseados em modelos**: estes algoritmos constroem um modelo estatístico que estima as preferências dos usuários. Eles podem usar redes Bayesianas, redes de dependência e diagnóstico de popularidade. O usuário então é classificado de acordo com uma dessas classes de modelos e seus interesses são pressupostos de acordo com esse modelo aprendido. É indicado para ambientes em que as preferências do usuário são pouco alteradas, porém, não recomendado para sistemas onde as preferências são alteradas frequentemente e rapidamente.

As técnicas de FC possuem alguns problemas bem conhecidos e analisados por (Sarwar et al. 2002), (Claypool et al., 1999), (Park et al., 2012). Um deles é o fato desta técnica não classificar as preferências de um usuário por sua própria opinião (seu perfil), mas estimá-la através de usuários que tem opiniões parecidas. No caso, a preferência de um usuário nem sempre reflete a de outro.

A FC sofre do problema de esparsidade, que se deve ao fato de que alguns itens acessados/adquiridos por um usuário podem ter tido pouco (ou talvez nenhum) acesso por outros usuários do sistema. Isso torna difícil a recomendação de itens similares a tais itens impopulares. Com isto, conteúdos mais populares, como o nome já diz, obtiveram maior número de acessos e, portanto, terão maiores chances de serem recomendados. Da mesma forma, itens pouco populares acabam sendo pouco (ou até nunca) recomendados. Assim, alguns usuários podem nunca ter acesso a tais conteúdos pela recomendação, devido à esparsidade de dados. Além disso, conforme descrito por (Balabanovic; Shoham, 1997), se um novo item aparecer na base de dados, não há como ele ser recomendado enquanto nenhum usuário avaliá-lo. Consequentemente, se um usuário tiver um gosto em particular muito diferente dos outros usuários no sistema, haverá chances de que ele não receba nenhuma recomendação.

Outro problema é a falta de escalabilidade da FC, que está relacionada com a complexidade dos algoritmos de filtragem de informação. Em muitos casos, a quantidade de informação (grande número de usuários e conteúdos/itens a recomendar) a ser levada em consideração no cálculo de uma recomendação é tão grande que o tempo de resposta pode acabar se tornando um problema.

Um dos maiores problemas ligados à FC é a chamada “partida a frio” (*cold-start)*, que dificulta a recomendação nos estágios iniciais do uso do sistema. Este problema é bem conhecido e foi analisado em trabalhos como (Burke, 2007), (Li; Kim, 2003), (Middleton et al., 2004), e muitos outros. O problema da partida a frio se deve ao fato de que inicialmente existem poucos usuários que tenham definido suas preferências e, portanto, o sistema de recomendação tem dificuldade em encontrar usuários com perfis semelhantes devido à falta de informação que se encontra inicialmente.

Um problema semelhante à partida a frio é o “problema do novo item”, descrito por (Pedronette; Torres, 2008). Este problema ocorre quando há a inclusão de um novo conteúdo no sistema. Esse novo conteúdo possui poucas ou nenhuma avaliação/acesso por parte dos usuários, dessa forma dificultando sua recomendação.

Em (Herlocker et al., 2012), os autores propõem uma técnica de FC para RDs onde o usuário especifica suas necessidades em forma de questão e o sistema recomenda itens baseado nas avaliações feitas por outros usuários com questões similares. Nesta proposta, quando o usuário faz uma consulta no sistema de busca, o sistema realiza dois tipos de buscas e apresenta o resultado dividido em duas partes: resultados da busca baseada no critério da busca definido pelo usuário, e uma lista com recomendações de conteúdos que tiveram melhores avaliações de outros usuários que utilizaram critérios de busca semelhantes. O grande diferencial dessa proposta está no fato da FC ser realizada com base não na comparação entre perfis de usuários, mas sim entre os itens resultantes das buscas realizadas por eles. Nesse sistema, o usuário, durante uma sessão, faz uma busca em forma de perguntas literais. Esta pergunta então é mantida no sistema e comparada com outras perguntas feitas anteriormente por outros usuários, com o intuito de encontrar perguntas semelhantes. Em outras palavras, não são perfis de usuários que são armazenados e comparados, mas questões feitas por eles. Para determinação do nível de relevância de um conteúdo, o usuário deve explicitamente indicar a relevância de cada item do resultado da busca. Esta é justamente uma das deficiências desta proposta, devido a avaliação explícita por parte do usuário a ser realizada.

Em (Cazella et al., 2008), os autores propõem uma técnica de recomendação de artigos acadêmicos, usando a abordagem de FC. Esta técnica utiliza a construção explícita do perfil do usuário. Na primeira parte, os usuários têm seus perfis modelados através de avaliações que eles realizaram sobre conteúdos. A partir do uso desses perfis, é realizado o cálculo da similaridade entre o perfil do usuário alvo e os outros perfis de usuários cadastrados. Para isso é usada a fórmula chamada "coeficiente de Pearson", muito utilizada na FC (Herlocker et al., 1999), (Breese et al.,1998), (WeiWei et al., 2009). Na segunda parte, através do resultado do cálculo da similaridade é identificado o subconjunto de usuários com maior grau de similaridade (os "vizinhos próximos"), que serão considerados no cálculo da predição do conteúdo a ser recomendado. Na terceira parte, são normalizadas as avaliações fornecidas pelos usuários sobre o documento em análise para recomendação e calcula-se a predição, com a média ponderada das avaliações dos "vizinhos próximos" com os respectivos pesos de similaridade. Nesta técnica, o próprio usuário deve explicitamente fornecer seu perfil através de avaliações sobre conteúdos do sistema,

portanto, esta técnica tem as mesmas limitações de (Herlocker et al., 2012).

Em (WeiWei et al., 2009), os autores propõem uma técnica de recomendação que parte do princípio de que um usuário que avaliou certo item com uma nota, provavelmente irá atribuir notas semelhantes a itens semelhantes. Essa proposta se baseia nas informações referentes aos atributos dos itens avaliados. Nela é construída uma matriz de avaliação (*m*, *n*) que relaciona os usuários *m* aos itens *n*. Para preencher a falta de informação devido à falta de avaliação de usuários a certos itens, o algoritmo busca as informações dos atributos (propriedades) dos itens avaliados por um usuário *m* e busca essas informações em outros itens, calculando-se e atribuindo uma nota a cada um desses itens semelhantes não avaliados, preenchendo assim os espaços vazios nessa matriz de avaliação. Essa etapa é realizada antes de se calcular os possíveis "vizinhos próximos". Em testes realizados foi constatado que essa técnica resultou em uma melhor precisão de recomendação mesmo para quando houve uma grande quantidade de dados sendo avaliados.

Sistemas de Recomendação vêm sendo cada vez mais explorados também no contexto educacional (Ferro et al., 2011) e (Sibaldo, 2007). O objetivo do uso de sistemas de recomendação neste contexto é de buscar e filtrar informações relevantes a um aluno ou professor em particular a partir de uma coleção geral de conteúdos educacionais.

(Primo et al., 2010) propõe um modelo de recomendação de conteúdos educacionais descritos por metadados. Tal modelo visa capturar as informações do usuário implicitamente, utilizando o padrão de metadados OBAA (Bez et al., 2009). A proposta mostra uma arquitetura interessante, que inclui elementos de Web Semântica no perfil dos usuários e nos conteúdos disponíveis. Porém é um modelo teórico, sem formalizar o procedimento realizado para a recomendação.

Em (Ferro et al., 2011), os autores fazem uso de um sistema híbrido de recomendação de materiais didáticos, que combina recomendação pela avaliação de conteúdos feitas por outros usuários, recomendação de conteúdos semelhantes ao perfil do usuário, e recomendação de conteúdos mais populares. Em tal proposta, o perfil do usuário é modelado de forma explícita, onde ele deve informar suas áreas de interesse.

Sistemas de recomendação não se enquadram apenas na Web. Na TV digital esses sistemas têm sido usados e melhorados cada vez mais. Conforme descrito em (Cremonesi; Turrin, 2010), os sistemas de recomendação têm sido cada vez mais úteis no domínio da TV Interativa (TVI), onde a presença de grandes catálogos de itens (como filmes,

shows de TV) drasticamente reduz a visibilidade de cada um deles. Os sistemas de recomendação mais usados para TVI são os de filtragem colaborativa, pois se baseiam na opinião geralmente implícita dos usuários, como em (Bernhaupt et al., 2008) e (Cremonesi; Turrin, 2010).

3.1.2 Filtragem Baseada em Conteúdo

As técnicas de Filtragem Baseada em Conteúdo (Boutemedjet; Ziou, 2008), (Mooney; Roy, 1999), a exemplo da FC, se baseiam na construção de um perfil para cada usuário, onde são definidas as preferências e interesses do usuário. Mas diferente da técnica de FC, a FBC se baseia na identificação de conteúdos relevantes ao usuário comparando o perfil do usuário com os metadados descritores do conteúdo ou com o próprio conteúdo de cada item da coleção. Portanto, a FBC trata-se de uma filtragem de informação que não considera os perfis dos outros usuários do sistema, apenas do usuário foco da recomendação.

De acordo com (Burke, 2007), a FBC gera recomendações a partir de duas fontes: as características associadas aos itens (seja seu próprio conteúdo, ou os metadados que o descrevem) e a avaliação dada aos itens pelo usuário (seja ela implícita ou explícita). Em (Middleton et al., 2004), os autores definem que a FBC observa o comportamento do usuário e recomenda novos itens que tenham relação com o perfil do usuário. (Balabanovic; Shoham, 1997) considera que um sistema de recomendação puramente construído na FBC é aquele em que as recomendações são feitas ao usuário baseando-se unicamente no perfil construído através da análise dos conteúdos dos itens que o usuário avaliou no passado.

Uma das vantagens da FBC, segundo (Mooney; Roy, 1999), diz respeito à privacidade de acesso a dados. Como na FC os itens são recomendados de acordo com os perfis de outros usuários semelhantes, um certo esboço do que outros usuários têm acessado fica evidente. Além disso, outra vantagem da FBC é o fato de que, diferente da FC, itens que foram recentemente introduzidos no sistema e que podem ser de interesse do usuário são recomendados numa FC. Isto ocorre porque a recomendação não requer a avaliação do novo item por outros usuários, ou seja, é feito apenas um cálculo de similaridade entre características do novo item e do perfil do usuário.

Em (Lopes et al., 2006), os autores propõem um sistema de recomendação para repositórios digitais de artigos científicos. Este sistema captura o perfil do usuário implicitamente a partir de uma

análise semântica do currículo Lattes do usuário. As informações como "título" e "palavras-chave" contidas no Lattes são usadas para determinar os termos do vetor de busca. Após determinado os vetores de busca é usada a abordagem tf X idf. Assim, os melhores termos para a identificação de um documento são aqueles com uma alta frequência de aparecimento em um documento (*tf*) e baixa frequência de aparecimento na coleção de documentos (alto valor de *idf*).

Em (Mooney; Roy, 1999) é proposto um sistema de recomendação baseado na FBC para recomendação de livros. Este sistema utiliza métodos automatizados para categorização de textos aplicados em textos semiestruturados extraídos da Web que descrevem os livros. Neste método, para a classificação do livro, são extraídas informações textuais de características que vão além das descritas por metadados. As informações utilizadas a respeito dos livros extraídos da web provêm do Amazon.com. Tais características constituem de trechos do conteúdo ou opiniões a respeito do livro. Neste sistema, existe uma construção explícita do perfil do usuário, onde os usuários fornecem notas de 1 a 10 para cada livro de um conjunto de treinamento. Estas notas servem como entrada para o treinamento para estabelecer o perfil do usuário usando um algoritmo de redes Bayseanas. Com base neste perfil, o sistema produz uma lista ordenada dos livros mais recomendados do catálogo. Para melhorar a recomendação, o usuário pode ainda avaliar os itens que lhe foram recomendados para dizer quais foram mal recomendados e exigir que o sistema refaça uma nova recomendação com melhores resultados.

Em (Martínez et al., 2007) é proposta uma técnica de recomendação para comércio eletrônico baseada na FBC. Nesta proposta, o processo de recomendação se dá por três fases: aquisição do perfil do usuário e das características dos itens; filtragem de itens, pela medição da similaridade entre o perfil do usuário e entre os itens armazenado na base; e construção da recomendação, ordenando os itens de acordo com seu grau de similaridade com o perfil do usuário. Nesta proposta, o perfil do usuário é construído explicitamente através da definição de um perfil onde o usuário, de forma subjetiva, informa suas necessidades de acordo com suas próprias percepções independentes dos conteúdos que procura. Esta técnica utiliza o conceito de contexto linguístico multigranular, onde as preferências e metadados dos itens serão modeladas. O uso de linguística multigranular permite que os usuários definam suas preferências em diferentes termos linguísticos de acordo com seu conhecimento, evitando que o esboço das preferências seja fornecido na mesma escala.

A FBC também sofre com a partida a frio. Para um usuário recém-cadastrado, o sistema de recomendação possui poucas informações de acesso para a modelagem de seu perfil. O sistema de recomendação terá uma estimativa dos interesses do usuário apenas após o usuário acessar ou avaliar certo conjunto de itens.

Outra desvantagem da FBC é o "fechamento dos itens a serem recomendados", ou também chamado de "estreitamento da análise de conteúdos" (Torres et al., 2004). Este problema se dá pela recomendação de itens similares em conteúdo com itens já avaliados ou acessados pelo usuário, impedindo a recomendação de itens que podem ter conteúdo diferente, porém relacionados. Por exemplo, para sites de comércio eletrônico com vários tipos de itens a serem oferecidos, um usuário que buscou por um livro de autoajuda, pode acabar por receber recomendações apenas de outros livros de autoajuda, desconsiderando outros materiais úteis como CDs, matérias ou artigos sobre autoajuda.

3.1.3 Técnicas híbridas

Apesar de muito difundidas e eficientes, as técnicas de FC e FBC também apresentam desvantagens uma em relação à outra. Por isso, para combinar as vantagens de ambas as técnicas, foram propostas algumas técnicas híbridas de recomendação (Basilico; Hofmann, 2004), (Torres et al., 2004), (Vellino; Zeber, 2007). Estas técnicas híbridas geram a recomendação baseando-se tanto no perfil do usuário que busca a recomendação (FBC), como no perfil de outros usuários similares (FC). De acordo com (Burke, 2007), sistemas híbridos tentam melhorar o desempenho, geralmente para lidar com o problema da partida a frio. Por sua vez, (Mooney; Roy, 1999) conclui que as FBC e FC são complementares, e que, juntando ambas as técnicas (abordagem híbrida), podem-se conseguir resultados mais satisfatórios para a recomendação.

Em (Balabanovic; Shoham, 1997), os autores apresentam o sistema de recomendação *Fab*, parte da biblioteca digital da Universidade de Stanford. A arquitetura desse sistema de recomendação possui três componentes principais: Agentes de coleta, que encontram páginas da Web que satisfazem determinado tópico; Agentes de seleção, que exploram as páginas identificadas pelos agentes de coleta e encontram páginas relevantes para um usuário em específico; e o roteador central. Cada agente gerencia um perfil: um agente de coleta gerencia determinados assuntos (determinados por palavras-chaves contidas nas páginas) em que as páginas identificadas da Web se situam

baseado em palavras-chave contidas nas páginas que foram localizadas; e um agente de seleção gerencia os interesses específicos de determinado usuário através das avaliações feitas por tais usuários das páginas recebidas. As páginas encaminhadas pelos agentes de coleta ao roteador central são direcionadas aos usuários cujos perfis se assemelham ao conteúdo da página recomendada (abordagem baseada na FBC). Após o recebimento de determinadas páginas relevantes, o usuário é solicitado para explicitamente avaliar seus conteúdos. Tais avaliações atualizam o perfil de seu agente de seleção. Com isso, quaisquer páginas que foram avaliadas positivamente serão recomendadas a outros usuários com perfis semelhantes (abordagem baseada na FC). Tais recomendações são processadas pelos agentes de seleção dos usuários que a recebem.

Em (Martins et al., 2007) é proposto um sistema de recomendação híbrido para RDs que objetiva a recomendação de artigos científicos baseando-se nas informações obtidas no currículo Lattes do usuário (construção implícita do perfil) e suas avaliações sobre os artigos acessados (captura explícita). O sistema é compatível com RDs que fazem uso de metadados no formato DC, com suporte ao protocolo OAI-PMH (*Open Archives Initiative*). Sua arquitetura conta com um conjunto de módulos: Módulo Lattes, que analisa o currículo Lattes do usuário cadastrado e o armazena na base de dados; o Módulo de Coleta OAI, que envia requisições aos provedores de dados em busca dos metadados no formato DC dos documentos cadastrados em cada provedor; Módulo de avaliação, onde o usuário avalia os artigos de seu interesse cadastrados no seu currículo Lattes; e o Módulo Dados Históricos, que armazena informações sobre os artigos já consultados pelo usuário. Dois modelos de recomendação híbridos foram testados pelos autores: O misto, e o ponderado. O primeiro executa a FC em paralelo com a FCB e apresenta a lista de resultados por ordem de relevância, que é dada pela soma dos escores obtidos pela FC e FCB. No segundo algoritmo, como descrito pelos autores, o peso de um determinado item recomendado é computado dos resultados de todas as técnicas de recomendação no sistema. Foram realizados testes de comparação entre algoritmos híbridos. A métrica utilizada, de forma geral, buscava determinar a quantidade de "acertos" nas recomendações se comparadas com o clássico *Top-n* (conteúdos mais recomendados). Os autores concluíram que o algoritmo misto, que literalmente junta os resultados da FC com FBC, apresentou melhores resultados.

O site Amazon (www.amazon.com) também utiliza uma técnica híbrida que considera outros compradores, e o perfil do usuário, que é

capturado via os itens comprados pelos usuário. Este sistema de compras usa uma técnica de FC para identificar usuários próximos, e a recomendação apresenta a lista de itens adquiridos por estes vizinhos próximos. O site Amazon também utiliza uma técnica de FBC que analisa o histórico de compras do usuário para recomendar, por exemplo, itens em lançamentos que são similares aos já adquiridos pelo usuário. Também faz uso da recomendação de itens sem a observação do perfil dos usuários, onde é recomendado itens ao usuário que foram comprados por outros usuários que compraram o item selecionado pelo usuário foco da recomendação (o popular, "quem comprou este item também comprou os seguintes itens"). Essa abordagem de recomendação de itens será visto na seção seguinte.

3.1.4 Outras Técnicas de Recomendação

As técnicas descritas anteriormente, apesar de serem as mais difundidas, não são as únicas existentes. Existem diversos trabalhos que propõem outras técnicas de recomendação que não as baseadas em FC e FBC, como Recomendação de Itens, Recomendação demográfica, e Recomendação baseada em conhecimento.

Na Recomendação de Itens, o perfil dos usuários não é essencial para a recomendação. Nesta técnica, os metadados dos itens são geralmente usados para a recomendação (Heylighen; Bollen, 2002). Nela, os itens disponíveis são recomendados de acordo com suas semelhanças com outros itens no sistema. Tais itens podem ter sido considerados semelhantes através da descrição de seus conteúdos ou metadados que os definem, ou por registros que mostram que tais itens tendem a ser comprados em conjunto com outros itens. Independente do perfil do usuário, ao acessar/avaliar um item, o usuário pode receber recomendações de outros itens devido à similaridade entre estes itens. Por exemplo, um usuário que compra uma vara de pescar, pode receber recomendações de iscas artificiais para compra. Esta é uma das técnicas bastante utilizadas principalmente em sites de comércio virtual em conjunto com as técnicas já citadas.

Uma das formas de se recomendar itens é através da coassociação. Dois conteúdos estão coassociados quando são acessados por um ou mais usuários, isto é, dois conteúdos *a* e *b* estão coassociados se um ou mais usuários que acessaram o conteúdo *a* também acessaram o conteúdo *b*. (Wasserman; Faust, 1994).

Em (Heylighen; Bollen, 2002), os autores propõem uma técnica de recomendação baseando-se num princípio semelhante ao da

coassociação, chamado pelo autor da *coativação*. Nessa técnica, os conteúdos acessados por um mesmo usuário recebem um nível de coativação uns com os outros, que é inversamente proporcional ao tempo que o usuário leva entre as consultas realizadas. Por exemplo, na coativação, se um usuário acessou os itens *a* e *b* num intervalo de tempo curto, e os itens *a* e *c* num intervalo de tempo um pouco mais longo, o item *a* terá um grau de coativação maior com o item *b* (ou seja, *a* estará mais relacionado com *b)* que com o item *c*. Dessa forma, após uma coleta de dados de tempos de acesso entre vários documentos por vários usuários o sistema acumula um grau de relação entre os documentos. Usuários seguintes, ao acessarem um documento, receberão recomendação de uma lista de documentos que tenham maiores relações com os que o usuário acessou recentemente. Além disso, seus acessos a partir daí também serão computados na relação entre os documentos acessados.

Outra técnica existente é a Recomendação Demográfica (Krulwich, 1997), onde as recomendações fornecidas são baseadas no perfil demográfico do usuário. Como descrito por (Burke, 2007), os produtos recomendados podem ter vindo de diferentes nichos demográficos. Neste tipo de recomendação são formados diferentes grupos de usuários que abrigam informações não apenas de suas preferências, mas especificamente quanto a informações pessoais do usuário como idade, gênero, posição geográfica, costumes, etc. Tais informações ajudam a definir o que certo tipo de usuário irá preferir. Em (Krulwich, 1997), é utilizado um sistema que divide um grupo de 40000 pessoas ao redor dos EUA em 62 grupos demográficos. Tal base demográfica contém mais de 600 variáveis que definem as características do estilo de vida de cada um. Fazendo uso dessa base, os dados de um usuário são comparados com os clusters para verificar quais características se assemelham e melhor generalizam o perfil de um usuário. No final, após a classificação a qual (ou quais) grupo(s) o usuário se classifica, conteúdos são recomendados de acordo com as características generalizadas que o definem. Tal técnica tem suas semelhanças a FC se considerar que os grupos modelados foram obtidos através da análise do perfil de outras pessoas (que na verdade não são usuários, mas dados demográficos obtidos anteriormente numa população).

Outra técnica de recomendação é a baseada em conhecimento, (Burke, 2007). Tal técnica leva em consideração que problemas similares possuem soluções similares (Schmitt; Bergmann, 1999). É voltada para usuários que não possuem muita informação sobre o

conteúdo que buscam. Portanto, através do uso de informações adicionais de buscas semelhantes, pode-se usar o conhecimento extra para trazer os conteúdos buscados pelo usuário em questão. Tais conhecimentos extras provêm de informações adicionais que foram incluídas em pesquisas semelhantes (porém mais ricas) feitas por outros usuários anteriormente. Por exemplo, um usuário buscando informações a respeito de um item sem saber o seu título, então ele fornece informações como formato, tamanho, e uso; e o sistema apresenta os resultados baseados no critério da busca. Porém, com o uso da técnica baseada em conhecimento, o sistema utiliza informações de critérios de busca anteriores semelhantes, porém mais detalhadas, geradas por outro usuário anteriormente e, assim, apresenta resultados mais específicos que condizem com o que o usuário inicial procurava.

Com o advento da Web 2.0, cada vez mais as tecnologias de web semântica também vêm sendo aplicadas na área de recomendação. Alguns trabalhos têm adotado estas tecnologias na construção de perfil ontológico para geração de recomendações. (Yu et al., 2005) define ontologia como uma descrição precisa de um conceito de domínio e é uma ótima forma de representação de conhecimento. A ontologia pode ser usada em várias etapas da recomendação, tanto na modelagem do perfil do usuário como na seleção dos itens a serem recomendados.

O perfil ontológico pode ser moldado com informações que vão além dos dados acessados por ele no sistema de recomendação. Informações como conhecimento do meio e áreas de interesse são também armazenados de forma estruturada e organizada. Uma vez que um domínio tenha sido classificado em termos de conceitos ontológicos, de acordo com (Middleton et al. 2009), as relações definidas pela ontologia do domínio podem ser usadas para inferir interesse e relevância de um conceito através de interesses que foram observados em outro conceito.

Um sistema baseado em conhecimento pode fazer uso de regras do sistema para inferir um interesse em classes de itens com uma conexão semântica para um item de interesse observado. Segundo experimentos apresentados em (Yu et al., 2005), a similaridade baseada no conceito é mais eficiente que a similaridade baseada em palavras-chave nos sistemas de recomendação personalizada. As técnicas de recomendação de itens que utilizam ontologias podem utilizar, por exemplo, ferramentas que exploram a semântica dos itens, a fim de evitar ambiguidades nos termos a serem procurados. Alguns frameworks usados incluem ODP, WordNet (Amini et al., 2011), DMOZ (Middleton et al., 2004), HowNet (Yu et al., 2005).

(Ruotsalo, 2010) afirma que os métodos baseados em ontologias podem ser usados para reduzir problemas que a FBC costuma apresentar. Esses problemas condizem com a forma como os sistemas analisam o conteúdo que recomendam (as características usadas para representar os objetos precisam ser extraídas automaticamente, pois a associação manual é uma tarefa muito trabalhosa); a forma com que trazem os conteúdos (é limitada pelas características explicitamente associadas com os objetos); e a forma como tratam a heterogeneidade dos conteúdos.

3.2 AGRUPAMENTO/CLUSTERIZAÇÃO EM RECOMENDAÇÃO

Uma das principais técnicas adotadas para reduzir o problema de escalabilidade na FC é o agrupamento (também chamado de clusterização) de usuários ou itens. O agrupamento visa reduzir quantidade de informações, incluindo características de itens ou perfis de usuários, sobre o qual a FC será realizada. O objetivo, portanto, é reduzir o tempo de processamento dos algoritmos de recomendação.

Apesar da vantagem da redução do custo computacional, as técnicas de agrupamento podem gerar certa perda de precisão dos resultados, pois a busca do sistema de recomendação é limitada a um subgrupo específico, excluindo outras informações (usuários ou conteúdos) (Sarwar et al., 2002).

Existem muitas técnicas de agrupamento, sendo que a *k-means* é a mais utilizada nos sistemas de recomendação (Sarwar et al., 2002), (Kim; Yang, 2005), (Li; Kim, 2003). Nessa técnica é definida inicialmente a quantidade *k* de clusters a serem criados e são escolhidos aleatoriamente *k* dados (perfis de usuários ou itens) do conjunto a ser comparado e os coloca como centroide de cada cluster. Em seguida, são distribuídos os demais dados nos clusters. Um dado é alocado a um cluster quando o dado é mais similar com o centroide do respectivo cluster. A cada ciclo, os centroides de cada cluster são atualizados (Jain; Dubes, 1988). Um cluster costuma ter uma abrangência de valores aceitáveis, onde cada elemento contido nele possui um valor que varia entre o mínimo e o máximo permitido no cluster. O "valor de centroide" é o valor central do cluster, onde seus elementos terão um valor mais próximo deste do que o do centroide de outros clusters. O valor de centroide pode ser atualizado como a média dos valores dos elementos que compõem o cluster.

No contexto de sistemas de recomendação, as técnicas de agrupamento são geralmente combinadas com a FC. (Sarwar et al., 2002) e (Kim; Yang, 2005) fazem uso do agrupamento para alocar usuários em grupos menores, objetivando reduzir o número de usuários a serem pesquisados na etapa de identificação dos vizinhos próximos.

Sarwar et al. (2002) propõem uma técnica que procura clusterizar a base de avaliações explícitas de conteúdos realizadas pelos usuários. Nela é usada uma variante da técnica *k-means*, onde se tem inicialmente um cluster contendo todos os dados. Em seguida, é calculado o valor do centroide desse cluster e então se faz uma bisseção neste ponto, separando o cluster inicial em dois subclusters. O processo continua recursivamente nos subclusters formados até se atingir o número *k* de clusters designado. Essa técnica gera clusters agrupados de forma hierárquica. Possui a vantagem de não gerar clusters vazios.

Kim e Yang (2005) propõem uma variante do *k-means* voltada a repositórios digitais em que, após a formação dos clusters, o algoritmo verifica se no cluster em que reside o usuário foco da recomendação há um número suficiente de vizinhos para gerar a recomendação. Em caso negativo, o algoritmo explora os clusters adjacentes a ele, isto é, aqueles que tenham seus valores de centroide mais próximos do cluster em questão, a fim de encontrar mais usuários vizinhos para o cálculo da recomendação.

A técnica proposta em (O'Connor; Herlocker, 2001) busca facilitar o processo de predição do resultado da avaliação de itens pelos usuários. Nela, algoritmos de clusterização são realizados no conjunto de itens, e a similaridade entre itens é usada para classificá-los e inseri-los nos clusters. Tais características dos itens a serem comparadas são baseadas nas avaliações que foram realizadas aos itens pelos usuários. Com os itens separados em clusters, cada cluster comporta tipos semelhantes de itens focados num único tópico, facilitando a predição de avaliações dos usuários aos itens não avaliados. Uma vez realizada a clusterização, uma técnica tradicional de FC é aplicada para recomendação de itens. Com o escopo da busca por itens similares é reduzida graças à clusterização, a FC aplicada sobre os itens terá uma dimensionalidade reduzida.

Em (Li; Kim, 2003) também é realizada uma clusterização de itens, porém com foco maior na resolução do problema da partida a frio. Os grupos de itens são formados a partir das avaliações feitas por grupos de usuários. Nesta técnica, itens novos que não foram avaliados ainda recebem suas predições de avaliação com base na média das avaliações

dos itens restantes do grupo. Dessa forma, o problema da partida a frio é reduzido, visto que os itens novos já recebem uma avaliação prévia.

3.3 CONSIDERAÇÕES FINAIS

Neste capítulo foram apresentadas várias propostas de Sistemas de Recomendação. Foram descritas várias técnicas e definições por parte de diferentes autores ao longo de quase 25 anos de estudos sobre o tema. Existem várias técnicas de sistemas de recomendação, cada uma voltada para um ambiente. Como parte do aprimoramento das ferramentas, alguns sistemas de recomendação foram implementados em conjunto com outras técnicas para melhor alcançar um objetivo, como a qualidade da recomendação ou tempo de processamento.

Há um grande número de repositórios digitais voltados a um certo tipo de conteúdo, seja ele artigos científicos, materiais didáticos ou, no caso da BD-LB, obras literárias. RDs fazem uso de metadados para melhor classificar as informações dos conteúdos disponíveis. Assim, foi desenvolvida uma técnica de recomendação que busca tirar vantagem das características de tais repositórios, buscando reduzir o custo computacional da recomendação sem interferir na qualidade da mesma.

O próximo capítulo apresenta detalhadamente a técnica de recomendação proposta.

4 TÉCNICA DE RECOMENDAÇÃO PROPOSTA

Este capítulo apresenta a técnica de recomendação proposta nesta dissertação. Esta técnica se baseia na Filtragem Colaborativa (FC) e agrupamento de usuários com perfis similares. Neste trabalho, esta técnica é analisada e avaliada em repositórios digitais (RDs). Apesar disto, esta técnica não é apenas exclusiva para RDs, mas para qualquer sistema onde os conteúdos disponibilizados são descritos por um conjunto de metadados através da modelagem de algumas regras descritas a seguir.

A técnica de recomendação proposta segue quatro etapas: construção dos perfis dos usuários; processo de agrupamento de usuários; determinação dos vizinhos próximos em cada grupo; e construção da recomendação (lista ordenada de conteúdos recomendados para o usuário). Ela assume a hipótese que os conteúdos acessados por grupos de usuários com perfis semelhantes possuem grande probabilidade de ser de interesse de um usuário individual pertencente a este grupo.

Graças à descrição dos conteúdos por valores de metadados, a construção dos perfis de usuários é realizada de uma maneira simples e eficiente, como proposto por (Willrich et al., 2006). O perfil do usuário é construído de maneira implícita, observando os valores dos metadados dos conteúdos acessados pelo usuário.

Com a preocupação da eficiência para repositórios digitais com grande número de usuários e coleções a serem analisados a fim de gerar a recomendação, foi adotada uma técnica de agrupamento de usuários. Na técnica proposta os grupos são criados a partir de uma lógica simples e eficiente. Os grupos são identificados por valores de metadados que são observados pelo sistema, por exemplo, grupos de determinado assunto, grupos de determinado autor. Um grupo é criado quando um usuário tem uma determinada frequência de ocorrência do valor de um metadado observado no conjunto de conteúdos já acessados. Por exemplo, se o usuário tem determinado número de acessos a obras de Machado de Assis, ele será incluído no cluster Machado de Assis. Se ele tem determinado número de acesso ao gênero Romance, ele será incluído no grupo Romance. Note que o usuário pode ser incluído em mais de um grupo. Cada grupo agrupa usuários com o mesmo interesse em determinado valor de metadado. Muitos modelos de clusterização para FC utilizam modelos onde cada usuário (ou conteúdo) possa ser agrupado em apenas um cluster. Na verdade, é mais natural assumir que

usuários (ou conteúdos) possam pertencer a múltiplos clusters (Xu et al., 2012).

O conjunto de usuários que são analisados para determinar os vizinhos próximos do usuário foco da recomendação é formado pelos membros dos grupos que este usuário faz parte. A técnica utiliza uma função clássica de cálculo de similaridade utilizando a fórmula do cosseno (Salton; Buckley, 1988) para determinar os vizinhos próximos deste conjunto. Finalmente, o sistema determina uma lista ordenada de conteúdos acessados pelos vizinhos próximos do usuário, retirando desta lista os conteúdos já acessados pelo próprio usuário. A ordenação é calculada com base em um cálculo do grau de preferência do conteúdo para o usuário.

4.1 PERFIL DO USUÁRIO

A técnica de recomendação proposta se baseia na associação de um perfil a cada usuário cadastrado no repositório digital, permitindo capturar as preferências e interesses de cada usuário. O modelo de perfil de usuário adotado é baseado no que foi proposto em (Furtado et al., 2009). Este modelo permite determinar o perfil do usuário de maneira implícita (sem intervenção do usuário), analisando a frequência de ocorrência dos valores de um subconjunto dos elementos de metadados que ocorrem nos conteúdos acessados pelo usuário.

Esta técnica se baseia na definição de Biblioteca Digital requerida por (Furtado et al., 2009) e é aqui definida como:

$$RD = (\mathrm{D}, \mathrm{M}, \mathrm{U}, \mathrm{PU}) \quad (1)$$

Em que:

- $D = \{d_i \mid i \in [1, I]\}$ define o conjunto de documentos que compõem a coleção do RD, com um total de I documentos.
- $M = \{md_j \mid j \in [1, J]\}$ define o conjunto de elementos de metadados utilizados que descrevem os documentos da coleção, onde J é o número de elementos de metadados. A notação $d_i.md_j$ é utilizada para referenciar o valor do elemento de metadado md_j do documento d_i.
- $U = \{u_k \mid k \in [1, K]\}$ é o conjunto de usuários do RD, onde K é o número de usuários registrados no sistema.
- $PU = \{p_k \mid k \in [1, K]\}$ é o conjunto de perfis de usuários. Cada usuário u_k é associado a um perfil p_k.

Para cada usuário u_k, o perfil p_k é definido pela fórmula (2). Na modelagem do perfil do usuário para a técnica de recomendação proposta, as DPUs e PGs não são consideradas. A cada usuário u_k é associado um pu_k, definido por:

$$pu_k = (PE_k, CA_k) \quad (2)$$

Onde:

- Preferências Específicas (PE) são as preferências do usuário relacionadas ao domínio do repositório. Elas são capturadas implicitamente por uma análise de frequência dos valores de um subconjunto dos elementos de metadados;
- Conteúdos Acessados (CA) são os conteúdos acessados pelo usuário. O acesso ao conteúdo sugere que o usuário se interessa neste tipo de conteúdo.

O conjunto de elementos de metadados utilizados para determinar as PEs é chamado de Metadados Observados (MO). A determinação dos MOs depende do domínio do repositório. Para tal, é necessário analisar que metadados são relevantes na determinação das preferências do usuário. A título de exemplo, na BD-LB observou-se que o autor e gênero literário são elementos de metadados (adotados na BD-LB) que são relevantes para determinar se uma obra é relevante ou não para o usuário.

Os MOs são definido por:

$$MO = \{mo_j \mid j \in [1, J]\} \quad (3)$$

- MO é um subconjunto dos elementos de metadados observados para a construção do perfil do usuário;
- J é o número de elementos de metadados observados.

O conjunto de documentos acessados por um usuário é definido por:

$$CA_i = \{d_k \mid k \in [0, K_j]\} \quad (4)$$

Onde d_k é um conjunto de documentos acessados pelo usuário i e K_j é o número total de conteúdos acessados por este usuário.

As preferências específicas de um usuário são estimadas com base em uma análise na frequência dos valores dos elementos de metadados observados. As PEs são definidas da seguinte forma por (Furtado et al., 2009):

$$PE_i = \{SV_j \mid j \in [1, J]\} \quad (5)$$

onde SV_j é um conjunto de valores ponderados do elemento de metadado observado mo_j construído a partir dos conteúdos acessados pelo usuário u_j. Por exemplo, SV_{autor} é a lista ponderada dos nomes dos autores cujas obras foram acessadas pelo usuário u_1. Neste conjunto, cada valor de um elemento mo_j está associado com um peso de preferência (w) que pode variar entre 0 e 1. Assim, SV_j é representado como um conjunto de valores ponderados:

$$SV_j = \{ vw_n, \mid n \in [1, N_j]\} \quad (6)$$

onde:

- N_j é o número de valores distintos do elemento de metadado mo_j encontrados nos conteúdos acessados pelo usuário u_i.
- vw_n é o valor ponderado (v_n, w_n), onde v_n é um valor do elemento de metadado mo_j em conteúdo acessado pelo usuário u_i e $w_n \in [0,1]$ representa o peso de preferência do valor de v_n do elemento de metadados mo_j. Quanto maior for esse peso, maior será o interesse do usuário para este valor de elemento específico.

Os conjuntos de valores ponderados dos elementos de metadado observados (SV) são atualizados quando um usuário acessa um novo documento. Quando o usuário acessa um documento d_i, cada valor do elemento de MO tem seu peso de preferência incrementado. Por exemplo, na BD-LB, quando um usuário acessa um conto escrito por José de Alencar, o peso de preferência do valor "José de Alencar" em SV_{autor} e o peso de preferência do valor "conto" em $SV_{gênero}$ devem ser incrementados.

Com base em (Chen et al., 2002), existem três análises básicas utilizadas para incrementar os pesos de preferência dos valores de MOs:

- **Análise de Existência**: o peso de preferência de um valor de MO pode assumir dois valores: 0 e 1. Caso o usuário não tenha acessado documentos com este valor de MO, o peso será 0, senão será 1. Note que neste modo a quantidade de acesso a documentos com o valor de MO não é considerado.
- **Análise de Frequência**: o peso de preferência associado com um valor de MO é definido pela frequência de ocorrência deste valos nos documentos acessados. Ou seja, é considerado o número de vezes que este valor ocorre nos documentos acessados. Existem ao menos dois métodos

possíveis para estimar o peso de preferência: *Whole-History* (WH), o peso de preferência é definido pelo número total de acessos; e o método *Past-Days* (PD), onde serão considerados apenas os valores dos MOs encontrados nos documentos acessados pelo usuário nos últimos *n* dias.

- **Análise de Idade de Acesso**: esta análise considera que os acessos mais recentes são mais relevantes para considerar a preferência do usuário do que os acessos mais antigos. Em (Chen et a., 2002), os autores propõem uma equação para calcular o peso baseado na idade acesso.

Nesta proposta, a técnica independe do tipo de análise utilizada para determinar o peso de preferência do valor do MO. Esta escolha dependerá do tipo de conteúdo oferecido pelo RD. Algumas considerações devem ser realizadas na escolha da análise:

- A análise de existência é muito simplista, e só poderia ser adotada em casos em que a simples ocorrência do valor de MO em um dos conteúdos acessados possa indicar que o usuário tem preferência por conteúdos com este valor. Por ser um valor binário, não é possível fazer diferenciações de preferência com base na frequência de ocorrência destes valores. Além disso, ela é válida apenas para situações em que as preferências do usuário não se alteram ao longo do tempo;
- A análise de frequência pode ser aplicada quando as preferências do usuário não mudam ao longo do tempo, ou são preferências que se mantêm por um longo período de tempo.
- A análise de idade de acesso é útil quando as preferências do usuário se alteram ao longo do tempo. Ou seja, ela é adequada quando se considera que acessos mais recentes devem ser considerados mais relevantes para estimar as preferências do que acessos mais antigos.

No caso do domínio da literatura, considerou-se empiricamente que as preferências do usuário não se alteram muito ao longo do tempo. Mas consideramos que testes devem ser realizados para verificar a validade desta tese. Devido a estas considerações, nesta dissertação foi adotada, a título de exemplo, o método WH para calcular o peso de preferência. Sendo assim:

$$w_n = numAcessos(mo_e, v_n)/numDocsAcessados(u_k) \quad (7)$$

Onde:

- numAcessos(mo_e,v_n) é o número de documentos acessados com o valor do elemento de metadados mo_e igual a v_n;
- numDocsAcessados(u_k) é o número total de documentos acessados pelo usuário u_k.

Para ilustrar os conceitos apresentados, considere o repositório BD-LB, que faz uso de dois *MO*s para a construção do perfil dos usuários (autor e gênero literário). De acordo com a equação (3) temos J=2, onde:

- mo_1 = Autor
- mo_2 = Gênero (literário)

Considere três usuários do repositório: u_1, u_2, e u_3. Considere também que cada um deles tenha realizado acesso a 10 documentos distintos da BD-LB, sendo que:

- u_1: acessou 7 obras de "Machado de Assis" e 3 obras de "José de Alencar", sendo que 8 destas obras são do gênero "Conto" e 2 obras são do gênero "Romance".
- u_2: acessou 4 obras de "Gregório de Matos" e 6 obras de "Machado de Assis", sendo que 5 destas obras são do gênero "Poesia" e 5 obras são do gênero "Conto".
- u_3: acessou 3 obras de "Machado de Assis", 2 obras de "Aluízio Azevedo", 3 obras de "José de Alencar" e 2 obras de "Gregório de Matos", sendo que 3 destas obras são do gênero "Conto", 2 obras são do gênero "Romance", 2 obras são "Poesia" e 3 são do gênero "Teatro".

De acordo com a equação (7) para determinação dos pesos de preferência dos valores do MOs, as PE destes usuários são:

- PE_1 = { {(Machado de Assis, 0.7), (José de Alencar, 0.3)} {(Conto, 0.8), (Romance, 0.2)} }
- PE_2 = { {(Gregório de Matos, 0.4), (Machado de Assis, 0.6)} {(Poesia, 0.5), (Conto, 0.5)} }
- PE_3 = { {(Machado de Assis, 0.3), (Aluízio Azevedo, 0.2), (José de Alencar, 0.3), (Gregório de Matos, 0.2)} } {(Conto, 0.3), (Romance, 0.2), (Poesia, 0.2), (Teatro, 0.3)} }

Os pesos de preferência associados aos valores dos elementos de metadados são atualizados sempre que um usuário acessa um novo conteúdo, usando a análise na frequência. Nesse momento, cada valor

dos elementos de MO tem seu peso de preferência incrementado. Por exemplo, se o usuário u_1 acessar um novo romance de "José de Alencar", suas PEs passariam para {{(Machado de Assis, 7/11), (José de Alencar, 4/11)} {(Conto, 8/11), (Romance, 3/11)}}.

4.2 PROCESSO DE AGRUPAMENTO

O processo de agrupamento se dá pela análise dos perfis dos usuários do RD. A determinação dos grupos a que um determinado usuário será adicionado depende dos pesos de preferência de determinados valores de MOs. Quando um determinado valor de MO atingir um limiar de peso de preferência, o usuário será incluído no grupo deste valor. Este limiar é chamado de *limiar de agrupamento*.

O conjunto de agrupamentos de usuários de um repositório (G) é definido por:

$$G = \{ g_{j,v_n} \mid j \in [1, \mathrm{J}], \mathrm{n} \in [1, \mathrm{N}_j] \} \qquad (8)$$

Sendo que:

- J é o número de MO considerados no agrupamento;
- N_j é o número de valores distintos do metadado j.

Na equação (8) observa-se que o número de grupos dependerá do número de MO considerados para aplicação da técnica de agrupamento e também do número de valores distintos destes MOs. Portanto, diferente do *k-means*, o número de grupos é dependente da alta frequência de diferentes valores de metadados dos perfis dos usuários. Quanto mais dispersos são os perfis, mais grupos serão criados.

O Algoritmo 1 especifica a técnica de agrupamento proposta. Este algoritmo é executado periodicamente (p.e. uma vez a cada 24h). Inicialmente o conjunto de grupo G é vazio. Em seguida, para cada perfil de usuário do RD é analisado o peso de cada um dos valores dos metadados observados. Caso o valor de um destes metadados atingir um limiar de agrupamento, o usuário será incluído no grupo identificado por este valor. Por exemplo, se o usuário realiza muitos acessos a obras de determinado autor ou palavra-chave, o usuário será incluído no grupo deste autor e no grupo desta palavra-chave.

Figura 1. Algoritmo de agrupamento adotado

```
Inicialização:
  G = {};
Para todos os usuários u_i do RD:
  Para todo SV_j ∈ PE_i
    Para todo vw_n ∈ SV_j
      Se w_n ≥ limiarAgrupamento_j
        Se grupo g_j,vn não existir, inclua este grupo em G
        Inclua o usuário u_i no grupo g_j,vn
```

O valor do limiar de agrupamento é específico do metadado observado. Seu valor depende principalmente da quantidade de valores distintos possíveis dos MOs. Por exemplo, os metadados observados da BD-LB são autores e gêneros literários. O número de autores possíveis atualmente é na ordem de milhares, enquanto os gêneros possíveis são 23 valores. Quando o elemento de metadado tem muitos valores possíveis, como autor, o limiar de agrupamento poderia ser menor, devido à possibilidade de maior dispersão dos usuários.

A título de ilustração, considere o MO de agrupamento Autor, os usuários u_1, u_2 e u_3 e um limiar comum de 0,4 (40%). A Tabela 1 reapresenta os pesos de preferência do MO Autor para os usuários considerados.

Tabela 2. Pesos de Preferência dos usuários u_1, u_2 e u_3.

Usuário	Autor	Preferência
u_1	Machado de Assis	0.7
u_1	José de Alencar	0.3
u_2	Gregório de Matos	0.4
u_2	Machado de Assis	0.6
u_3	Machado de Assis	0.3
u_3	Aluízio Azevedo	0.2
u_3	José de Alencar	0.3
u_3	Gregório de Matos	0.2

Considerando o limiar de agrupamento de 40%, serão criados os grupos do metadado Autor apresentados na Tabela 2. Estes grupos foram determinados aplicando o algoritmo de agrupamento (1). Note que o usuário u_3 não será incluído em nenhum grupo. Considera-se que o perfil de u_3 não permite ainda determinar as preferências deste usuário,

e nenhuma recomendação será feita a este. Ou opcionalmente, podem-se recomendar conteúdos utilizando outras técnicas mais simples, como as obras mais populares, ou ainda, realizar a recomendação normal sem a etapa de agrupamento

Tabela 3. Grupos de MO Autor e seus usuários.

Grupo Autor	Usuários
Machado de Assis	u_1, u_2
Gregório de Matos	u_2

Na nossa técnica, a lógica simples de agrupamento e criação de grupos não padroniza o número de grupos ou similaridade de tamanho. Assim, é possível obter grupos com apenas 1 usuário, e grupos com vários usuários.

A vantagem da técnica de agrupamento proposta se dá, além do processamento reduzido do algoritmo de recomendação aplicado a um grupo, pelo fato de que usuários com tais características de alta frequência de ocorrência de valores de MOs têm maior grau de similaridade entre eles. Dessa forma, o algoritmo de FC, ao explorar apenas os usuários de determinado grupo, tenderá a encontrar usuários com maior grau de similaridade entre si, sem a necessidade de explorar outros usuários que tenderão a ter pouca (ou nenhuma) relação com os usuários do grupo pesquisado. Assim, pode-se beneficiar do custo reduzido de processamento sem prejudicar a qualidade da recomendação.

4.3 DETERMINAÇÃO DOS VIZINHOS PRÓXIMOS

Nessa terceira etapa, o objetivo é identificar os "vizinhos próximos" de um usuário Q. Para tal, é calculada a similaridade entre o perfil do usuário Q e o perfil de cada um dos outros usuários que pertencem aos grupos que o usuário Q pertence. Este cálculo é realizado em duas etapas. Na primeira é determinada a similaridade entre os conjuntos de valores ponderados de cada um dos valores dos metadados observados (*SV*s). A segunda etapa determina a similaridade de perfis de usuário.

Para o cálculo de similaridade entre *SVs* de um metadado observado foi adotada a função do cosseno. Muitas técnicas de FC fazem uso de cálculos como o coeficiente de Pearson para determinar

um grau de similaridade entre usuários. O coeficiente de Pearson é usado para dar uma correlação entre duas variáveis (que podem ser usuários, documentos, etc, dependendo de como for aplicado) e fornece um resultado variando entre +1 e -1. Nela, +1 significa total correlação, 0 nenhuma correlação e -1 total correlação negativa (onde a tendência de uma variável indica falta da tendência de outra). No repositório em questão, na etapa de busca por usuários similares, buscou-se encontrar similaridades entre perfis de usuários sem haver uma correlação negativa. Dessa forma, adotamos a fórmula do cosseno para calcular a similaridade, que fornece resultados entre 0 (nenhuma relação entre os perfis) e +1 (perfis idênticos). Portanto, a similaridade entre os conjuntos SVs de um metadado observado mo_i é determinada por:

$$similaridade_{\,mo_i}(Q,D) = \frac{\sum_{n=1}^{N} w_{qn}.w_{dn}}{\sqrt{\sum_{n=1}^{N}(w_{qn})^2.\sum_{n=1}^{N}(w_{dn})^2}} \qquad (9)$$

Onde:

- Q representa o usuário foco da recomendação;
- D representa outro usuário do cluster a ser comparado;
- N é o número de valores diferentes do elemento de metadado mo_i definidos no conjunto SV_i do perfil do usuário Q.
- w_{qn} é o peso de preferência do valor v_{in} do elemento de metadados mo_i obtido do perfil do usuário Q.
- w_{dn} é o peso de preferência do valor de v_{in} do elemento de metadados mo_i se presente no perfil do usuário D. Senão, w_{dn} é zero.

A título de exemplo, considere a recomendação sendo realizada para o u_1 apresentado anteriormente. Os usuários u_1 e u_2 terão seus perfis comparados, pois ambos estão contidos nos grupos “Machado de Assis” e “Conto”. Nesta etapa é calculada a similaridade dos metadados observados autor e gênero:

- $similaridade_{\,mo_{autor}}(1,2) = 0.76$
- $similaridade_{\,mo_{gênero}}(1,2) = 0.69$

A segunda etapa é o cálculo da similaridade entre perfis de usuários. Este último é obtido pela média simples entre as similaridades de cada MO da equação (9), conforme equação (10):

$$similaridade\ _{perfil}(Q,D) = \frac{\sum_{i=1}^{I} similaridade_{mo_i}}{I} \quad (10)$$

Para os usuários do exemplo, o cálculo de similaridade dos perfis resultaria em:

- $similaridade\ _{perfil}(1,2) = 0.73$

Vale lembrar que, com a adoção do agrupamento dos usuários, u_1 e u_3 não serão comparados por não possuírem nenhum grupo em comum, visto que u_3 não foi incluído nem no grupo "Machado de Assis" nem em "Conto".

Nesta proposta, os vizinhos próximos de um usuário Q são todos aqueles que têm uma similaridade de perfil com o usuário Q acima de um valor, chamado aqui de limiar de similaridade, como mostrado na fórmula (11). Este valor de limiar é dependente do domínio do repositório e dos metadados observados selecionados. No caso da BD-LB, após a realização de testes preliminares, foi definido empiricamente o valor de 0,7 como limiar inicial.

$$Vizinhos_Q = \{D | similaridade_{perfil}(Q,D) \geq limiar\} \quad (11)$$

Como pôde ser visto no exemplo recém-explicado, e considerando um limiar de similaridade de 0.7, o usuário 2 é o único vizinho próximo do usuário 1.

4.4 CONSTRUÇÃO DA RECOMENDAÇÃO

Esta próxima etapa tem por objetivo produzir a lista ordenada de conteúdos a serem recomendados a um determinado usuário. Esta lista é obtida a partir do conjunto de conteúdos acessados pelos vizinhos próximos do usuário foco da recomendação, ou seja, que tenham uma similaridade de perfil acima de um limiar.

Esta técnica assume que a lista de recomendação deve ser constituída de conteúdos que não foram ainda acessados pelo usuário. Desta forma, os conteúdos já acessados pelo usuário serão retirados da lista de conteúdos recomendados.

A lista de recomendação é ordenada com base no cálculo do peso de preferência definido por (Willrich et al., 2006). Desta forma, os conteúdos que tiverem maior chance de serem interessantes para o

usuário devem aparecer primeiro. A lista ordenada de conteúdos recomendados *CR* para um usuário u_k é definida como segue:

$$CR = (d_1, d_2, \ldots, d_R) \text{ onde } \forall\, r \in [1, R-1], \quad w_{pref}(u_k, d_r) \geq w_{pref}(u_k, d_{r+1})). \tag{12}$$

O peso de preferência wd_r de um conteúdo d_r a ser recomendado é estimado pela combinação dos pesos de preferência de cada um dos valores dos elementos de metadado observados deste conteúdo. Para exemplificar o cálculo do peso de preferência, considere novamente o usuário u_1 da BD-LB com seu perfil descrito na seção 4.1, e considere a Tabela 3 abaixo como sendo a lista de conteúdos a serem recomendados para u_1:

Tabela 4. Lista de Recomendação para o usuário 1.

Título	Autor	Gênero Literário
Niterói	José de Alencar	Poesia
Casa velha	Machado de Assis	Romance
A paixão	Manuel de Paiva	Conto

De acordo com as PEs de u_1 (apresentado na seção 4.1), o valor de preferência de cada conteúdo será calculado como a média entre o peso de preferência dos valores dos elementos de metadado "autor" e "gênero". O resultado deste cálculo é apresentado na tabela 4.

Tabela 5. Cálculo do peso de preferência dos conteúdos.

Título	w_{autor}	w_{genero}	W_{pref}
Niterói	0,3	0	0,15
Casa velha	0,7	0,2	0,45
A paixão	0	0,8	0,4

Finalmente, com base no peso de preferência calculado, a lista de conteúdo pode ser ordenada, conforme apresentado na Tabela 5.

Tabela 6. Lista Ordenada de Itens Recomendados.

Título	wd
Casa velha	0,45
A paixão	0,4
Niterói	0,15

4.5 CONSIDERAÇÕES FINAIS

A proposta descrita busca tirar vantagem das características de estrutura de RDs com escopo de conteúdos textuais como na BD-LB. Através do desenvolvimento de tais técnicas ordenadas passo a passo, é possível conciliar o uso dos metadados adotados com um sistema de recomendação que facilita a busca de informações em tais RDs.

Na nossa proposta é aplicada uma técnica de agrupamento de usuários com base nas maiores frequências de acesso de conteúdos. O algoritmo possui a vantagem de sua simplicidade e fácil consulta e ordenação para a definição dos grupos a serem criados. No algoritmo de agrupamento proposto não ocorrem cálculos recursivos de separação de cluster, nem predefinições de quantidade ou tamanho de clusters como no *k-means*. A seleção de usuários a dividirem um mesmo grupo se dá por uma consulta numa tabela de dados que guarda as maiores frequências de metadados acessados pelos usuários. O uso dessa técnica busca agrupar usuários que possuem grandes frequências nos mesmos valores de metadados.

Na nossa técnica, com o uso do cálculo do cosseno, usuários com grandes frequências de acesso a um mesmo valor de metadado têm grandes chances de obterem um alto valor de grau de similaridade. Assim, o uso da técnica de agrupamento proposta visa limitar a busca por usuários com perfis similares (melhorando o desempenho da recomendação), bem como explorar apenas os usuários que possuem maiores chances de terem um valor de similaridade maior, consequentemente excluindo da busca aqueles que tiverem maiores chances de terem um grau de similaridade menor, ou irrelevante (sem afetar a qualidade da recomendação).

Para a avaliação desta proposta, foram realizados experimentos a fim de testar a sua eficiência. Os testes buscam mostrar a qualidade da recomendação proposta para usuários reais, bem como simulações mostrando o comportamento da recomendação em um ambiente com grande quantidade de usuários cadastrados no sistema.

A seção seguinte mostra a realização destes experimentos, descrevendo cada etapa da execução na BD-LB, bem como a conclusão dos resultados obtidos.

5 IMPLEMENTAÇÃO E TESTES REALIZADOS

A seguir é descrita a implementação da técnica de recomendação proposta na BD-LB. Além disso, são descritos os resultados dos testes realizados.

5.1 IMPLEMENTAÇÃO NA BD-LB

A BD-LB possibilita aos usuários o cadastro de uma conta, onde ele informa, entre outros, seu nome, endereço e email. Sua conta estará em seguida associada ao perfil do usuário, capturado implicitamente como definido na seção 4.1.

No contexto deste trabalho, o código da BD-LB foi alterado de modo a implementar a técnica de recomendação proposta neste trabalho. A recomendação de obras é realizada na página principal da BD-LB, após o usuário realizar a autenticação.

5.1.1 Metadados Observados

A BD-LB conta com uma grande quantidade de obras literárias disponíveis. Para melhorar a classificação de seu conteúdo e a modelagem do perfil dos usuários, foram adotados dois metadados: "Autor" e "Gênero Literário".

Os conteúdos da BD-LB possuem outros dados que também os descrevem, como: século, ano de publicação, ou idioma. Para a modelagem do perfil, foram considerados apenas os metadados Autor e Gênero por serem considerados os mais importantes na descrição de obras de literatura brasileira. Fato este também comprovado em uma pequena pesquisa de opinião com alunos do curso de letras da UFSC e outros usuários da BD-LB.

5.1.2 Parametrização do Agrupamento

O sistema de recomendação apresenta a lista de obras recomendadas utilizando o agrupamento por frequência de acessos de 0,4 tanto para o metadado autor, quanto para o metadado gênero. Isso significa que qualquer perfil de usuário que obtiver frequência de acessos a um valor de metadados >= a 40%, será incluído no grupo daquele valor de metadados. Através de testes preliminares foi concluído que tal valor foi o melhor a ser adotado no ambiente em que foi implementado. Valores muito baixos fizeram com que vários grupos

de usuários pouco similares fossem criados, enquanto que valores muito altos fizeram com que pouquíssimos grupos fossem criados.

Tal valor pode ser ajustado também de acordo com a quantidade de diferentes valores dos metadados adotados. Por exemplo: para o valor de metadado Gênero, há poucas variações, podendo ser adotado um valor de limiar de frequência de acessos maior; enquanto que o valor de metadado Autor tem dezenas de milhares de diferentes valores, podendo ser adotado um valor de limiar de frequência menor, decido sua grande abrangência.

5.1.3 Parametrização da Determinação dos Vizinhos Próximos

Após o agrupamento, a FC usada determina a lista de usuários semelhantes do grupo com um limiar de similaridade de 0,7. Valores muito altos desse limiar fazem com quem poucos (ou nenhum) usuários relevantes sejam encontrados. De forma similar, valores muito baixos fazem com que muitos usuários sejam considerados relevantes na FC, reduzindo o propósito da técnica de recomendação, que é de buscar uma quantidade de vizinhos próximos relevantes. Com base nos experimentos realizados, limitou-se o uso desse valor para a apresentação de uma recomendação que tente explorar uma quantidade mínima de usuários mais relevantes para se obter uma melhor qualidade de recomendação.

5.1.4 Construção da recomendação

Nesta implementação foram adotadas duas formas de recomendação. A primeira recomendação é baseada na popularidade e, portanto, composta das obras mais acessadas da base. Esta recomendação é apresentada a todos os usuários (mesmo não cadastrados) e independe do perfil de quem visualiza a página Web. Esta lista de recomendação por popularidade é apresentada na esquerda da Figura 2.

Figura 2. Recomendação de obras mais populares.

Quando o usuário se autenticar na BD-LB, ao lado da lista de recomendação por popularidade (Figura 2) é apresentada a lista de obras recomendadas ao usuário com base na técnica proposta neste trabalho, como mostrado na Figura 3.

Figura 3. Recomendação baseada no agrupamento e FC da proposta.

A Figura 3 apresenta a lista de recomendação para um dado usuário, seguindo o seu perfil. Da forma que as obras a serem recomendadas estão ordenadas pelo grau de preferência do usuário foco da recomendação. Inicialmente a lista apresenta os 5 resultados mais relevantes, sendo possível clicar no link “Veja Mais” para que se receba mais recomendações.

O usuário pode se interessar por uma ou mais obras listadas e acessar seu conteúdo, fazendo com que o seu perfil seja atualizado. Com essa mudança no perfil, se o usuário indo novamente a página inicial da BD-LB receberá uma nova lista de recomendação (o grau de similaridade com outros usuários pode sofrer uma mudança fazendo com que o sistema mude algumas obras na lista de recomendações).

5.1.5 Demais aspectos de implementação

Todo o processo de FC é realizado no momento em que o usuário cadastrado acessa a página principal. A etapa de agrupamento que antecede a Filtragem Colaborativa é organizada em determinados horários do dia para reduzir o processamento no momento da recomendação. Isso evita que o sistema refaça todo o processo de agrupamento a cada vez que um usuário peça por uma recomendação. A desvantagem de se realizar o agrupamento apenas em determinados horários se dá pelo fato de que o usuário ao buscar por uma recomendação pode estar recebendo uma baseada em grupos que não estejam totalmente atualizados.

Para a realização do agrupamento foram criadas duas novas tabelas no banco de dados da BD-LB, cada uma correspondente a um dos metadados adotados “Autor” e “Gênero Literário”. Cada uma delas contém apenas 2 colunas: uma contendo o id do usuário e outra contendo o id do valor do metadado em que ele está inserido. Tais valores vão sendo inseridos nas tabelas sempre que um perfil ultrapassar a frequência de acessos de 40% a um valor de metadados. Esses dados de frequência estão em outras tabelas que contém o id de cada usuário que acessou obras literárias junto com as respectivas frequências de valores de metadados. Essa etapa de agrupamento e atualização de frequências de acesso é um processo trabalhoso para ser realizado a todo instante, de modo que ela é realizada num determinado horário do dia, todos os dias. Esta abordagem possui a vantagem de economia de tempo no processo de agrupamento, evitando grandes demoras na recomendação; porém possui a desvantagem de que os dados de agrupamento não estão sempre com a versão mais atualizada, sendo que

uma recomendação pode ser dada com base em um agrupamento realizado horas antes.

5.2 TESTES REALIZADOS

Dois tipos de avaliação do sistema proposto foram realizadas. O primeiro visou avaliar a precisão da técnica de recomendação proposta. A segunda visou avaliar o impacto do agrupamento no desempenho do sistema.

5.2.1 Teste de precisão

As principais métricas de validação de resultados de sistemas de recomendação são precisão e revocação. A técnica de precisão é medida pela razão entre o número de documentos relevantes recuperados e o número total de documentos recuperados pelo sistema. Já a revocação é definida pela relação entre o número de documentos relevantes recuperados e o número total de documentos relevantes contidos no sistema. Para avaliar a técnica de recomendação proposta, foi utilizada a métrica de precisão. A métrica de revocação não foi possível ser realizada devido ao grande número de obras disponíveis para que os usuários avaliassem a relevância de cada obra disponível na BD-LB.

Ao todo, participaram deste teste 28 usuários reais da BD-LB. Inicialmente os participantes acessaram 5 obras de sua preferência. Em seguida, foi informado para cada usuário uma lista de 5 obras recomendadas. Os participantes então avaliaram cada obra com notas de 1 a 5 de acordo com seu nível de relevância para leitura, isto é, seus possíveis interesses de leitura por cada obra recomendada. Dos 28 participantes que se cadastraram e acessaram obras, 24 avaliaram as recomendações recebidas.

Para fins de cálculo da precisão, considerou-se que o conteúdo é relevante se a nota dada pelo usuário é superior ou igual a 3. Além disso, devido à baixa quantidade de usuários, adotou-se o limiar de similaridade de 0,5.

É importante levar em conta o contexto em que se aplicam os testes. Para a realização dos testes de precisão com usuários reais é necessário que tais usuários tenham algum conhecimento prévio a respeito dos conteúdos a serem avaliados, de forma que estejam capacitados de avaliar tais conteúdos. Como contra exemplo, poderíamos citar um sistema que recomenda as melhores traduções de textos para determinada língua estrangeira, e tentar validá-lo com

usuários que não dominem tal língua. Para evitar tais problemas, o grupo de usuários selecionados teve seu foco voltado às alunos dos cursos de graduação e pós-graduação em Literatura da UFSC.

A tabela 6 mostra as notas dadas para as recomendações recebidas, junto com a média e o índice de precisão.

Tabela 7. Avaliações dos usuários às recomendações recebidas.

Usuário	Obra 1	Obra 2	Obra 3	Obra 4	Obra 5	Média	Precisão
U1	4	1	1	5	4	3	0,6
U2	4	3	2	3	3	3	0,8
U3	2	5	5	5	5	4,4	0,8
U4	5	5	4	3	3	4	1
U5	5	4	5	5	3	4,4	1
U6	5	2	1	1	N.D.	2,25	0,2
U7	4	5	5	4	3	4,2	1
U8	5	5	4	4	4	4,4	1
U9	2	5	5	2	2	3,2	0,4
U10	4	5	5	5	5	4,8	1
U11	4	3	5	5	3	4	1
U12	3	4	3	2	5	3,4	0,8
U13	4	5	4	4	5	4,4	1
U14	3	3	3	4	3	3,2	1
U15	4	4	1	4	5	3,6	0,8
U16	2	5	5	3	3	3,6	0,8
U17	2	3	2	2	3	2,4	0,4
U18	3	5	5	5	4	4,4	1
U19	3	1	1	1	1	1,4	0,2
U20	2	5	5	3	1	3,2	0,6
U21	4	5	2	3	5	3,8	0,8
U22	5	5	5	5	4	4,8	1
U23	5	1	4	5	4	3,8	0,8
U24	2	5	5	3	5	4	0,8

A precisão descrita na tabela acima indica de forma percentual a quantidade de notas que ultrapassaram o parâmetro estabelecido para

serem consideradas válidas. O índice de precisão médio obtido neste experimento foi de 0,78 com desvio padrão de 0,25.

Apesar do reduzido número de usuários, este valor demonstrou a boa qualidade dos conteúdos recomendados. Este índice tende a aumentar de acordo com o aumento do número de usuários. Para verificar isso, foi realizado anteriormente um teste similar com apenas 6 usuários, onde a precisão média foi de 0,625 com desvio padrão de 0,14.

5.2.2 Tempo de Processamento

Esta seção apresenta os testes de tempo de processamento da técnica proposta, visando medir a escalabilidade do sistema proposto. Para a realização deste teste foram inseridos 10000 usuários fictícios no repositório e, para cada um deles, foi simulado o acesso a 10 obras dentre um conjunto de 748 obras.

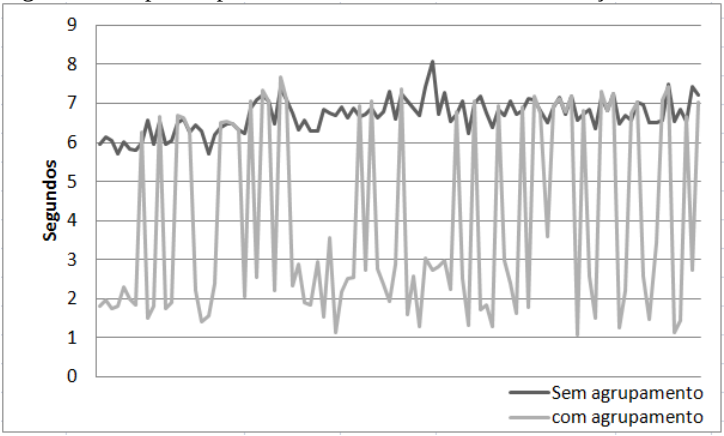
Para a seleção das 10 obras acessadas pelos usuários, tentou-se emular o comportamento dos usuários reais. Dessa forma, tentou-se simular a existência de autores populares e também levado em conta que geralmente usuários tendem a ler um pequeno conjunto de autores. Sendo assim, os autores mais populares, obviamente, são acessados por mais usuários. Para emular esse padrão, o acesso a 10 obras simulado pelo teste não foi totalmente aleatório: Para cada usuário, as seis primeiras obras foram limitadas a dois autores escolhidos aleatoriamente pertencentes ao grupo dos 15 autores mais populares do repositório, sendo 3 acessos para cada um dos dois autores. Em seguida, os quatro últimos acessos são aleatórios dentre todo o conjunto de obras digitalizadas.

Para viabilizar a realização do teste, foram escolhidos aleatoriamente 100 dos 10000 usuários cujos tempos de processamento das recomendações seriam medidos. Para cada um destes usuários foram geradas duas listas de recomendação, cada uma contendo as 10 obras mais recomendadas. Uma das listas foi gerada sem o uso da técnica de agrupamento e outra com o uso de agrupamento com limiar de agrupamento de 0,4. Para fins de comparação, foram medidos os tempos de processamento e também as listas de obras recomendadas. Para estes testes, foi utilizado um servidor não dedicado com processador Intel Quad-Core Xeon E5405 2GHZ, com 8GB de memória.

A Figura 4 apresenta os tempos de processamento médio das recomendações com e sem agrupamento. Em média, o tempo de processamento da recomendação sem agrupamento foi de 6,68s, com desvio padrão de 0,43s. Com a utilização da técnica de agrupamento

proposta obteve-se uma média de 3,9s, com desvio padrão de 2,37s. Estes resultados mostram que houve um ganho significativo no tempo de processamento com o uso da técnica de agrupamento. Deve-se notar aqui o elevado desvio padrão com o uso do agrupamento, que se deve ao fato de que alguns usuários não foram incluídos em nenhum grupo (por não alcançarem frequência de acesso superior ao limiar de 0,4). Para tais casos, a FC sem agrupamento é realizada, sem prejudicar o tempo de processamento, já que o tempo necessário para a recomendação será igual ao pior caso, como mostra a Figura 4.

Figura 4. Tempos de processamento médio das recomendações.



Ao todo, dos usuários que foram testados, 63 foram alocados em grupos, contra 37 que não foram. Analisando os resultados apenas dos que pertenciam a grupos, foi obtida uma média de tempo de recomendação de 2,13s, com desvio padrão de 0,62s. No total, foram criados 23 grupos (15 de autores, e 8 de gêneros), sendo o maior contendo 2521 usuários e o menor com 8 usuários. A média de tamanho dos grupos foi de 414. Como o teste selecionou aleatoriamente as obras, não há um comportamento natural dos usuários de selecionarem um menor número de autores e gêneros literários.

O processo de agrupamento visa reduzir o espaço de busca da FC a fim de obter um tempo de resposta otimizado. Uma característica negativa deste processo se dá pela possível redução da qualidade da recomendação.

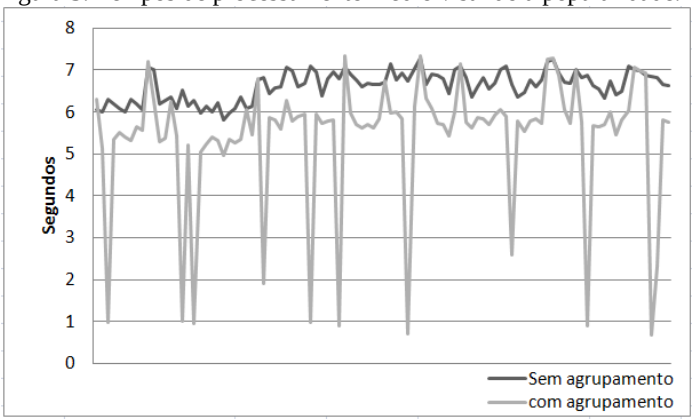
Para avaliar o impacto na qualidade da recomendação da técnica de agrupamento, foi realizado um estudo comparativo de qualidade entre duas listas de recomendação usando a técnica proposta,

sendo uma com a utilização da técnica de agrupamento e outra sem. Tais listas foram comparadas entre si para que pudesse ser avaliado o percentual de perda de qualidade entre as recomendações. Em todas as recomendações propostas para este teste específico as listas de recomendação geradas com e sem a técnica de recomendação eram idênticas. Portanto, neste teste realizado, a aplicação da técnica de agrupamento proposta não resultou em nenhuma perda de qualidade. Porém, isto não significa que a técnica é livre de perdas de qualidade, como será descrito mais adiante.

Tal resultado pode ser justificado devido ao cálculo do cosseno da FC utilizada que define os vizinhos próximos. Nele, usuários que tenham altas frequências (> 40%) de acesso ao mesmo valor de metadado terão alto valor no grau de similaridade entre si. Isto faz com que a grande maioria dos vizinhos próximos de um determinado usuário esteja situada nos mesmos grupos que o usuário em questão esteja alocado, tornando desnecessária a busca por vizinhos próximos em toda a lista de usuários cadastrados no sistema.

Outros testes foram realizados que consideravam outros critérios. Por exemplo, um outro teste realizado visava simular usuários reais partido da premissa que usuários tendem a acessar os conteúdos mais populares. Para isso, o acesso às 10 obras a ser simulado para cada usuário era realizado da seguinte forma: as 5 primeiras obras eram restritas às obram dos 10 autores mais populares (com mais acessos) da BD-LB, limitando o escopo a 437 obras. As 5 obras restantes seriam então dentre todas as 748 obras disponíveis. A Figura 5 mostra o resultado de tal teste. Em média, o tempo de processamento da recomendação sem agrupamento foi de 6,62s, com desvio padrão de 0,76s; enquanto a recomendação com agrupamento obteve uma média de 5,4s, com desvio padrão de 1,57s.

Figura 5. Tempos de processamento médio visando a popularidade.



Vários outros testes foram realizados, e em alguns deles foram detectadas perdas na qualidade da recomendação. Porém, tais perdas sempre se limitavam a no máximo 5 dos 100 pares de listas de recomendações comparadas. Além disso, dentre estas listas comparadas que estavam diferentes, tais diferenças se limitavam sempre entre as 2 ou 3 últimas obras de cada lista. Tal resultado, mesmo no pior dos casos testados, mostrou que o sistema proposto tem uma perda de qualidade insignificante perante ao ganho no tempo de processamento alcançado.

6 CONCLUSÕES

Este trabalho propôs uma técnica de recomendação de conteúdos para repositórios digitais, considerando a existência de metadados que descrevem os conteúdos. O objetivo desta técnica é reduzir o esforço despendido pelo usuário na localização de conteúdos relevantes. Este benefício foi comprovado via uma avaliação empírica.

A técnica proposta analisa valores dos elementos de metadado dos conteúdos acessados pelos usuários a fim de estimar as suas preferências, agrupá-los corretamente, e formular a recomendação. O objetivo é determinar as preferências do usuário a partir dos dados que descrevem os conteúdos acessados anteriormente. O conjunto de elementos de metadado a ser observado é dependente do domínio do repositório, permitindo que metadados descritores de objetos de aprendizagem sejam considerados.

Fazendo uso do registro de valores de metadados para modelar o perfil de preferências dos usuários, foi possível criar uma técnica de agrupamento simples e de baixo custo computacional para auxiliar no processo de recomendação baseado na FC. Tal técnica de agrupamento desenvolvida se destaca frente às outras pela análise dos registros de frequências de acesso dos usuários aos valores de metadados adotados e, em seguida, agrupa os usuários que possuírem alto valor de frequência de acessos aos mesmos valores de metadados. Como o processo explora toda a base de dados, seria dispendioso o uso de processamento se tal etapa fosse realizada a cada recomendação feita. Ao invés disso, ele é realizado num horário específico durante o dia, independentemente da quantidade de recomendações solicitadas no sistema. Como a modificação dos grupos não é constante, e uma sensível diferença não traria grandes impactos num sistema com grande quantidade de usuários, não há a necessidade constante de atualização do processo de agrupamento. Dessa forma, o agrupamento programado permite uma boa atualização dos grupos de usuários a serem alocados na base de dados.

Utilizando a alta frequenta de acessos a metadados como critério de alocação de usuários, a técnica permitiu que os usuários identificados como ‘vizinhos próximos’ pela FC estejam, em grande maioria, alocados no mesmo grupo em que o usuário foco da recomendação se encontra. Com isso, a técnica permitiu que a busca por usuários na FC limitasse seu escopo aos usuários com maiores chances de terem um perfil semelhante ao do usuário foco da recomendação.

Assim, foi possível criar um sistema de recomendação para RDs com menor custo computacional e sem perda da qualidade da recomendação.

Os resultados dos testes realizados mostraram que o sistema funcionou corretamente e atendeu as expectativas. O teste de stress com usuários simulados mostrou que houve uma redução significativa no tempo de processamento da recomendação com uma perda de precisão dos resultados insignificante (comparada à FC tradicional). Além disso, o teste com usuários reais mostrou que há uma boa aceitação do sistema proposto, que tende a ser cada vez mais eficiente com o aumento da quantidade de usuários cadastrados.

Como trabalhos futuros, pretende-se melhorar o algoritmo, incluindo o tratamento de casos especiais em que usuários não são alocados em nenhum grupo, além da implementação do agrupamento considerando mais de um metadado. Por exemplo, um grupo que aloca usuários com alta frequência para os valores de metadados "Machado de Assis" E "Contos". Outra proposta para trabalhos futuros seria a utilização do agrupamento por contexto, onde usuários cadastrados em um RD que tenha grande abrangência de conteúdos possam ser agrupados com relação ao conhecimento e interesse por determinados contextos disponíveis no RD. Por exemplo, para um RD que abrange os conhecimentos de uma universidade, os usuários poderiam ser agrupados por contexto do estudo, como biologia, matemática, etc. Além disso, pretende-se fazer uso das ferramentas de Web 2.0, com a inclusão de Ontologias e Web Semântica para o cálculo de conteúdos a serem recomendados.

Este trabalho resultou em uma publicação científica entitulada "**Técnica de Recomendação Baseada em Metadados para Repositórios Digitais Voltados ao Ensino**" no SBIE 2013 - XXIV Simpósio Brasileiro de Informática na Educação. Este evento possui Estrato Qualis/CAPES: B2. Tal publicação recebeu o primeiro lugar no prêmio de melhores artigos do SBIE 2013. Em virtude disso, houve ainda um convite para a submissão de artigos premiados na RBIE - Revista Brasileira de Informática na Educação, que será submetido até a metade de Abril de 2014.

REFERÊNCIAS

Amini, B., Ibrahim, R. et al. **Incorporating Scholar's Background Knowledge into Recommender System for Digital Libraries**. Em: 5th Malaysian Conference in Software Engineering (MySEC), p. 516-523. 2011.

Balabanovic, M., Shoham, Y. **Fab: Content-Based, Collaborative Recommendation**. Comm. ACM, vol. 40, no. 3, pp. 66-72, 1997.

Basilico, J., Hofmann, T. **Unifying Collaborative and Content-Based Filtering**. Em: Proceedings of the International Conference on Machine Learning, Banff, Alberta, ACM, New York, New York, 9-16. 2004.

Basu, C., Hirsh, H., Cohen, W. **Recommendation as Classification: Using Social and Content-Based Information in Recommendation**. Em: 15th National Conference on Artificial Intelligence, 714-720, 1998.

Bernardo, F. Turrin, R. Cremonesi, P. Ghsi, B. Siqueira, F.A. **SuggesTv: A Content Recommendation Application for Digital Television**. XVII Simpósio Brasileiro de Sistemas Multimídia e Web 2, 75-55, 2011.

Bernhaupt , R., Wilfinger, D., Weiss, A., Tscheligi, M. **An Ethnographic Study on Recommendations in the Living Room: Implications for the Design of iTV Recommender Systems**, Proceedings of the 6th European conference on Changing Television Environments, July 03-04, 2008.

Bez, M. R., da Silva, J. M. C., Santos, E. R., Primo, T. e Bordignon, A. **Projeto obaa: Uma abordagem com objetos de aprendizagem interoperáveis baseados na web e na televisão digital**. Informática na Educação: Teoria e Prática, 12(1), p. 119–126, 2009.

Boutemedjet, S., Ziou, D. **A graphical model for context-aware visual content recommendation**. IEEE Transactions on Multimedia, 10(1), p. 52–62, 2008.

Breese, J. S., Heckerman, D., Kadie, C. **Empirical analysis of predictive algorithms for collaborative filtering**. Em: 14th

Conference on Uncertainty in Artificial Intelligence (UAI ’98), 43-52, 1998.

Burke, R. **Hybrid Web Recommender Systems**. Em: Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.) The Adaptive Web: Methods and Strategies of Web Personalization. LNCS, vol. 4321, pp. 377–408. Springer, Heidelberg 2007.

Casagrande, M. F. R., Kozima, G., Willrich, R. **Técnica de Recomendação Baseada em Metadados para Repositórios Digitais Voltados ao Ensino**. Em: XXIV Simpósio Brasileiro de Informática na Educação, 2013. Disponível em: <http://www.br-ie.org/pub/index.php/sbie/article/view/2546>. Acesso em Jan. 2014.

Cazella, S. C., Chagas, I. C., Barbosa, J. L., Reategui, E. B. **Um Modelo Para Recomendação de Artigos Acadêmicos Baseado em Filtragem Colaborativa Aplicado à Ambientes Móveis**. RENOTE. Revista Novas Tecnologias na Educação 7, 12-22, 2008.

Cazella, Sílvio C., Nunes, M. A. S. N., Reategui, E. B. **A Ciência da Opinião: Estado da Arte em Sistemas de Recomendação**”. Em: Jornada de Atualização de Informática – JAI 2010 – CSBC. Rio de Janeiro: PUC RIO, v. 1, p. 161-216. 2010.

Chen, M., LaPaugh, A.S., Singh, J.P. **Predicting Category Accesses for a User in a Structured Information Space**. SIGRI’02, 2002.

Claypool, M. et al., **Combining Content-Based and Collaborative Filters in an Online Newspaper**. Proc. ACM SIGIR ’99 Workshop Recommender Systems: Algorithms and Evaluation, Ago. 1999.

Cleveland, G. **Digital libraries: definitions, issues and challenges**. IFLA Universal Dataflow and Telecommunications Core Programme, Occasional Paper 8, 1998. Disponível em: <www.ifla.org/VI/5/op/udtop8/udtop8.htm>. Acesso em nov. 2013.

Cremonesi, P., Turrin, R. **Time-evolution of IPTV recommender systems**. Eighth International Interactive Conference on Interactive TV & Video, pp. 105–114, 2010.

Crow, R. 2002. **The Case for Institutional Repositories: A SPARC Position Paper**. Scholarly Publishing & Academic Resources Coalition, Washington, D.C., 2002. Disponível em: <http://scholarship.utm.edu/20/1/SPARC_102.pdf>. Acesso em nov. 2013.

DanEr, C. **The collaborative filtering recommendation algorithm based on BP neural networks**. Em: Proceedings of the International Symposium on Intelligent Ubiquitous Computing and Education, p. 234–236, 2009.

Dublin Core Metadata Initiative (2014) Dublin Core Metadata Element Set, Versão 1.1. Disponível em: <http://dublincore.org/documents/dces/>. Acesso em março. 2014.

Duncan, C. **Digital Repositories: e-Learning for Everyone**. Presented at eLearnInternational, Edinburgh 9-12 Feb. 2003.

Ferro, M.R.C., Nascimento Júnior, H.M., Paraguaçu, F., Costa, E.B., Monteiro, L.A.L. **Um Modelo de Sistema de Recomendação de Materiais Didáticos para Ambientes Virtuais de Aprendizagem**. Simpósio Brasileiro de Informática na Educação, p. 810-819, 2011.

Furtado, C.A., Willrich, R, Fileto, R. Siqueira, F.L., Tazi, S. **Ordenação Personalizada na Recuperação de Informações em Bibliotecas Digitais**. Em: Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia), 2009.

Heery, R., Anderson, S. **Digital repositories review**. Report by UKOLN and AHDS. 2005. Disponível em: <http://www.jisc.ac.uk/uploaded_documents/digital-repositories-review-2005.pdf> . Acessado em Nov. 2013.

Heylighen F., Bollen J. **Hebbian Algorithms for a Digital Library Recommendation System.** Em: Proc. 2002 Int. Conf. on Parallel Processing Workshops (IEEE Computer Society Press), 2002.

Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J. **An Algorithmic Framework for Performing Collaborative Filtering**. Proc. 22nd Ann. Int',l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ',99), 1999.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., Riedl, J. T. **Evaluating collaborative filtering recommender systems**. ACM Trans. on Information Systems 22, 1 (Jan. 2004), 5–53. Disponível em <http://dx.doi.org/10.1145/963770.963772 > . Acesso em nov. 2013

Herlocker, J., Jung, S., Webster, J. G. **Collaborative Filtering for Digital Libraries**. Technical Report. Oregon State University. 2012. Disponível em: <http://hdl.handle.net/1957/28103>. Acesso em nov. 2013.

Huang, Z., Li, X., Chen, H. **Link prediction approach to collaborative filtering**. Em: Proceedings of the 5th ACM/IEEECS joint conference on Digital libraries (ACM Press, New York, 2005). Disponível em <http://dl.acm.org/citation.cfm?id=1065415>. Acesso em nov. 2013.

IEEE LTSC (2004): **IEEE LTSC Working Group 12: Learning Object Metadata**. Disponível em: <http://ltsc.ieee.org/wg12/index.html>. Acesso em nov. 2013.

Jain, A. K., Dubes, R. C. **Algorithms for Clustering Data**. Prentice-Hall, Inc., Upper Saddle River, NJ, 1988.

Kim, T-H., Yang, S-B. **An Effective Recommendation Algorithm for Clustering-Based Recommender Systems**. Em: Advances in artificial Intelligence, p. 1150-1153, 2005.

Krulwich, B. **Lifestyle Finder: Intelligent User Profiling Using Large-Scale Demographic Data**. Artificial Intelligence Magazine 18(2), 37-45, 1997.

Li, Q., Kim, B.M. **Clustering Approach for Hybrid Recommender System**. IEEE /WIC International Conference on Web Intelligence (WI'03), p. 33-38, 2003.

Lopes, G. R., et al. **Sistema de Recomendação para Bibliotecas Digitais sob a Perspectiva da Web Semântica**. Em: II Workshop de Bibliotecas Digitais. XXI Simpósio Brasileiro de Banco de Dados, 2006.

Maleki-Dizaji, S., Othman, Z.A., Nyongesa, H.O., Siddiqi, J. **Evolutionary Reinforcement of User Models in An Adaptive Search**

Engine. IEEE/WIC International Conference on Web Intelligence (WI 2003), pp. 706-709, 2003.

Martins, H. N. J., Costa, E. B., Oliveira, T. T. M., Bittencourt, I. I. **Sistema de Recomendação Híbrido para Bibliotecas Digitais que Suportam o Protocolo OAI-PMH**. Em: Anais do XXII Simpósio Brasileiro de Informática na Educação, SBIE, Aracaju, 2011.

Martínez, L., Pérez, L. G., Barranco, M. **A multigranular linguistic content-based recommendation model**. Research Articles, International Journal of Intelligent Systems, v.22 n.5, p.419-434, 2007.

Middleton, S.E., Shadbolt, N.R., Roure, D.C. **Ontological User Profiling in Recommender Systems**, ACM Trans. Information Systems, vol. 22, no. 1, pp. 54-88, 2004.

Middleton S.E., Roure D.D., Shadbolt N.R. **Ontology-based recommender systems**. Em: Staab, S., Studer, R. (eds) Handbook on Ontologies, International Handbooks on Information Systems, p. 779–796. Springer, Heidelberg 2009.

Middleton, S.E., Shadbolt, N.R., Roure, D.C. **Ontological User Profiling in Recommender Systems**. ACM Trans. Information Systems, vol. 22, no. 1, pp. 54-88, 2004.

Mooney, R.L., Roy, L. **Content-Based Book Recommending Using Learning for Text Categorization**. Proc. ACM SIGIR ',99 Workshop Recommender Systems: Algorithms and Evaluation, 1999.

Open Archives Initiative (OAI). Disponível em <http://www.openarchives.org/>. Acesso em Jan. 2014.

O'Connor, M., Herlocker, J. **Clustering Items for Collaborative Filtering**. Em: Proceedings of SIGIR-2001 Workshop on Recommender Systems, New Orleans, LA, 2001.

O'Donovan, J., Smyth, B., Gretarsson, B., Bostandjiev, S., Höllerer, T. **PeerChooser: visual interactive recommendation**. Proceedings of the 26th SIGCHI Conference on Human Factors in Computing Systems (CHI'08), ACM Press, New York, NY, USA. p. 1085–1088, 2008.

Park D.H., Kim H.K., Choi I.Y., Kim J.K. **A literature review and classification of recommender systems research**. Expert Systems with Applications, 39 (11), p. 10059-10072, 2012.

Pedronette, D., Torres, R. S. **Uma Plataforma de Serviços de Recomendação para Bibliotecas Digitais**. In: XXIII SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD), Campinas, São Paulo, 2008.

Primo, T.T., Vicari, R.M., Silva, J. M. C. **Rumo ao Uso de Metadados Educacionais em Sistemas de Recomendação**. Simpósio Brasileiro de Informática na Educação, pp. 4-8, 2010.

Resnick, P., Varian, H. R. **Recommender Systems**. Communications of the ACM 40, 3, p. 55-58. 1997.

Ruotsalo, T. **Methods and applications for ontology-based recommender systems**. Ph.D. Thesis, Aalto University, School of Science and Technology (Espoo, Finland), 2010.

Salton G., Buckley, C. **Term-weighting approaches in automatic text retrieval**. Information Processing and Management: an International Journal, v.24 n.5, p.513-523, 1988.

Santos, J. R., Ferraz, C. **Uma Arquitetura Flexível de Suporte à Execução de Algoritmos de Recomendação de Programas em TV Conectada**. XVII Simpósio Brasileiro de Sistemas Multimídia e Web 1, 97-102, 2011.

Sarwar, B., Karypis, G., Konstan, J., Riedl, J. **Analysis of recommendation algorithms for e-commerce**. Em: Proceedings of the 2nd ACM conference on Electronic commerce, p.158-167, 2000.

Sarwar, B., Karypis, G., Konstan, J., Riedl, J. **Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering**. Em: Fifth International Conference on Computer and Information Technology, p. 158–167, 2002.

Schmitt, S., Bergmann, R. **Applying case-based reasoning technology for product selection and customization in electronic commerce**

environments. 12th Bled Electronic Commerce Conference. Bled, Slovenia, June 7-9, 1999.

Sibaldo, M. A. A., Sales, T. B. M., Calado, I.A.A.R., Bittencourt, I.I., Costa, E.B. **Mobile GraW: Uma Aplicação para Dispositivos Móveis Baseada em Comunidades Virtuais de Aprendizagem Com Suporte A Recomendação**. Simpósio Brasileiro de Informática na Educação, p. 214-217, 2007.

Speretta, M., Gauch, S. **Personalized search based on user search histories**. IEEE/WIC/ACM International Conference on Web Intelligence, pp 622- 628, 2005.

Torres, R., McNee, S.M., Abel, M., Konstan, J.A., Riedl, J. **Enhancing digital libraries with techlens**. Proceedings of the 2004 joint ACMIEEE conference on Digital libraries(JCDL'04), p.228-228, 2004.

Vellino, A., Zeber, D. **A Hybrid, Multi-dimensional Recommender for Journal Articles in a Scientific Digital Library**. Em: Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, p.111-114, 2007.

Wan, G., Liu, Z. **Content-based information retrieval and digital libraries**. Information Technology & Libraries 27, 41–47, 2008.

Wasserman, S., Faust, K. **Social Network Analysis**. Cambridge University Press, 1994.

Weiwei, X., Liang, H., Junzhong, G., Keqin, H. **Effective collaborative filtering approaches based on missing data imputation**, NCM 2009-5th International Joint Conference on INC, IMS, and IDC , art. no. 5331661, pp. 534-537, 2009.

Willrich, R., Speroni, R.M., Lima, C.V., Diaz, A.L.O., Penedo, S.M. **Sistema de Recuperação de Informações Adaptativo Aplicado a Bibliotecas Digitais**. Simpósio Brasileiro de Sistemas Multimídia e Web, pp. 165-173, 2006.

Xu, B., Bu, J., et al. **An Exploration of Improving Collaborative Recommender Systems via User-Item Subgroups**. Proceedings of the 21st international conference on WWW, p 21-30, 2012.

Yu, Z., Zheng, Z., Gao, S., Guo, J. **Personalized information recommendation in digital library domain based on ontology**. Em: IEEE International Symposium on Communications and Information Technology, 2005. ISCIT 2005, vol. 2, p. 1249–1252, 2005.