

UM SERVIÇO PARA RECOMENDAÇÃO DE ARTIGOS CIENTÍFICOS BASEADO EM FILTRAGEM DE CONTEÚDO APLICADO A DISPOSITIVOS MÓVEIS

Silvio César Cazella *
Jonas Vinícius Drumm**
Jorge Luis V. Barbosa***

Resumo. Com a grande quantidade de informações disponíveis hoje em dia, e o fácil acesso às mesmas, através da Internet, a busca por informações relevantes tornou-se uma tarefa cada vez mais complexa. Neste trabalho, propõe-se a utilização de Sistemas de Recomendação na educação, visando a recomendação de conteúdo relevante aos alunos através do uso de dispositivos móveis. A vantagem de se utilizar sistemas de recomendação em dispositivos móveis está na facilidade da entrega da recomendação aos alunos. Com base neste cenário, este artigo apresenta um serviço de Sistema de Recomendação focado em Filtragem Baseada em Conteúdo (FBC) para ambientes móveis. O modelo proposto foi implementado em um protótipo visando disponibilizar um serviço para recomendação de artigos científicos. A avaliação das recomendações recebidas foi realizada pelos usuários utilizando uma escala likert de 5 pontos. No final deste artigo são descritos alguns experimentos realizados e os resultados obtidos com a aplicação do serviço proposto são discutidos.

Palavras-chaves: filtragem baseada em conteúdo, sistemas de recomendação, computação móvel.

A SERVICE FOR RECOMMENDATION OF PAPERS BASED ON CONTENT FILTERING APPLIED TO MOBILE DEVICES

Abstract. Due to the large amount of information available today, and the easy access to the Internet, the search for relevant information has become an increasingly complex task. We propose the use of recommendation systems in education, aiming at the recommendation of relevant content on mobile devices. The advantage to using recommendation systems on mobile devices is the ease of delivery of the content to users. Based on this scenario, this paper presents a service of recommendation system focused on Content-Based Filtering (CBF) to mobile environments. The model was implemented in a prototype aimed at the recommendation of scientific papers. The evaluation of the recommendations received by users was conducted using a 5 point Likert scale. At the end of this paper some experiments are described and the results obtained with the application of the proposed service are discussed.

Keywords: content based filtering, recommender systems, mobile computing.

1. Introdução

Com a grande quantidade de informações disponível hoje em dia, e o fácil acesso as mesmas através da Internet, as pessoas necessitam de ferramentas que agilizem o acesso

* Doutor em Computação. Professor Adjunto do Programa de pós-graduação em Engenharia da Produção, Universidade do Vale do Rio dos Sinos, Av. Unisinos 950, CEP 93.022-000, São Leopoldo, RS. E-mail: cazella@unisinos.br

** Bacharel em Ciência da Computação da Universidade do Vale do Rio dos Sinos, Av. Unisinos 950, CEP 93.022-000, São Leopoldo, RS. E-mail: jonasdrumm@gmail.com

*** Doutor em Computação. Professor Adjunto do Programa de pós-graduação em Computação Aplicada Unisinos, Universidade do Vale do Rio dos Sinos, Av. Unisinos 950, CEP 93.022-000, São Leopoldo, RS. E-mail: jbarbosa@unisinos.br

as informações efetivamente relevantes. Na maioria das vezes, o usuário não é capaz de selecionar as informações mais adequadas a sua necessidade, seja por desconhecimento do tema, ou pela sobrecarga de informação provocada pelas ferramentas de busca utilizada (exemplo, google¹).

Os sistemas de recomendação têm por objetivo auxiliar na obtenção da informação relevante ao usuário, seja pelo uso do seu perfil ou pelo perfil do grupo (Cazella et al., 2010). Os Sistemas de Recomendação que utilizam a técnica de filtragem denominada de filtragem baseada em conteúdo realizam uma filtragem de itens a serem recomendados baseando-se na análise do conteúdo do item já consumido no passado, o item a ser recomendado (por exemplo, uma notícia, um livro, um artigo científico) e o perfil do usuário, visando verificar se o item é ou não relevante para o usuário alvo da recomendação (Adomavicius et al., 2005).

Este artigo visa abordar os conceitos relativos a sistemas de recomendação que utilizam a Filtragem Baseada em Conteúdo (FBC), visando agregar essa funcionalidade ao projeto LOCAL (descrito em detalhes na seção 3). O novo serviço fornecerá ao módulo tutor do projeto LOCAL, recomendações de itens para dispositivos móveis baseado em FBC. Essa integração auxiliará no aprendizado dos estudantes com a aplicação de mobilidade computacional., A principal contribuição desta pesquisa esta na união entre sistemas de recomendação que utilizam filtragem baseada em conteúdo (envolvendo a complexidade de se lidar com textos), e a entrega destas recomendações através de dispositivos móveis ao corpo discente de um curso de graduação.

Este artigo está organizado em sete seções. Na seção 2, o conceito de Sistemas de Recomendação será apresentado, bem como a técnica de filtragem baseada em conteúdo (FBC). A seção 3 apresenta o projeto LOCAL com suas funcionalidades atuais. A seção 4 descreve o modelo que viabiliza o serviço proposto, bem como a arquitetura do mesmo. Na seção 5, os experimentos de avaliação do protótipo desenvolvido são descritos e os resultados discutidos.

2. Sistemas de Recomendação

Os sistemas de recomendação são fundamentalmente colaborativos, uma vez que permitem aos usuários compartilhar informações sobre materiais consultados/adquiridos, possibilitando também que registrem suas impressões sobre estes materiais. Estes sistemas têm assumido um importante papel no ambiente web no que diz respeito a auxiliar usuários no acesso a informação relevante ou na escolha de itens de seu interesse. Nesse contexto itens podem ser artigos científicos (Cazella et al., 2010); (Silva e Cazella, 2005), filmes (Schafer et al., 1999), músicas, livros, software ou produtos de consumo em geral.

Nos Sistemas de Recomendação são utilizadas em geral uma das três técnicas de filtragem de informação citadas a seguir: filtragem baseada em conteúdo (FBC), filtragem colaborativa (FC), também conhecida como filtragem social e filtragem híbrida (FH) (Cazella et al., 2010). Neste artigo será descrita em detalhe a técnica de Filtragem Baseada em Conteúdo devido a esta ter sido utilizada no modelo desenvolvido para viabilizar o serviço.

2.1. Filtragem Baseada em Conteúdo

Segundo Herlocker (Herlocker et al., 2000), por muitos anos os cientistas têm direcionado seus esforços para aliviar o problema ocasionado com a sobrecarga de informações através de projetos que integram tecnologias que automaticamente

¹ www.google.com

reconhecem e categorizam as informações. Alguns *softwares* têm como objetivo gerar de forma automática descrições dos conteúdos dos itens e comparar estas descrições com os interesses dos usuários visando verificar se o item é ou não relevante para cada um (Balabanovic e Shoham, 1997). Esta técnica é denominada de filtragem baseada em conteúdo (Herlocker et al., 2000), por realizar uma seleção baseada na análise de conteúdo dos itens e no perfil do usuário alvo da recomendação.

A abordagem baseada em conteúdo tem suas raízes na área de recuperação de informação. Devido aos significativos avanços feitos pelas comunidades de filtragem de informação e filtragem de conteúdo, muitos sistemas baseados em filtragem de conteúdo focam na recomendação de itens com informações textuais, como documentos e *websites*. As melhorias sobre os sistemas tradicionais de recuperação de informação vieram com a utilização do perfil do usuário, que contém suas preferências e necessidades.

As informações sobre o perfil do usuário podem ser obtidas do próprio usuário, como uma consulta realizada por ele, ou coletadas através do conteúdo dos itens que o usuário consome. Uma técnica muito comumente utilizada neste tipo de abordagem é a indexação de frequência de termos (*term frequency indexing*). Neste tipo de indexação, informações dos documentos e necessidades dos usuários são descritas por vetores com uma dimensão para cada palavra que ocorre na base de dados. Cada componente do vetor é a frequência que a respectiva palavra ocorre em um documento ou na consulta do usuário. Sendo assim, os vetores dos documentos que estão próximos aos vetores de consulta são considerados os mais relevantes para ele (Herlocker et al., 2000).

Desta forma, a filtragem baseada em conteúdo parte do princípio de que os usuários tendem a interessar-se por itens similares aos que demonstraram interesse no passado, definindo então, a similaridade entre os itens (Herlocker et al., 2000). Em alguns casos, pode haver maior dificuldade para estabelecer esta similaridade. Para que seja estabelecida a similaridade entre itens como roupas e brinquedos, por exemplo, seria necessária a identificação dos atributos nos itens a serem comparados (peso, preço, marca, etc). No caso dos itens serem artigos científicos (ou outros documentos) este processo de comparação pode ser facilitado, pois documentos podem ser considerados similares se compartilharem termos em comum. Sendo assim, a filtragem baseada em conteúdo é mais indicada para a recomendação de itens textuais, onde o conteúdo é geralmente descrito com *keywords* (Salton e McGill 1983). Um exemplo é o sistema de recomendação FAB (Balabanovic e Shoham 1997), o qual recomenda páginas a usuários. Este sistema representa uma página web com os seus 100 mais importantes termos. Conforme Adomavicius (Adomavicius et al., 2005), pode-se formalizar esta definição estabelecendo *ContentBasedProfile(c)* como sendo o perfil do usuário *c*. Este perfil é obtido através de uma análise do conteúdo dos itens previamente avaliados pelo usuário utilizando técnicas de recuperação de informação. Por exemplo, *ContentBasedProfile(c)* pode ser definido como um vetor de pesos (w_{c_1}, \dots, w_{c_k}) onde cada peso w_{c_i} denota a importância do termo k_i para o usuário *c* utilizando-se a medida TF-IDF (*term frequency-inverse document frequency*). Em sistemas de recomendação baseados em conteúdo, a função utilidade $u(c,s)$ é geralmente definida conforme a equação (1):

$$u(c,s) = \text{score}(\text{ContentBasedProfile}(c), \text{Content}(s)) \quad (1)$$

Tanto o *ContentBasedProfile(c)* do usuário *c* como o *Content(s)* podem ser representados como vetores (TF-IDF) de pesos e termos \vec{w}_c e \vec{w}_s . Além disso, a função utilidade $u(c,s)$ normalmente é representada, na literatura de recuperação de informação,

por algum tipo de pontuação heurística sobre vetores, como por exemplo, a medida de similaridade do cosseno. O cálculo para a medida de similaridade do cosseno pode ser feito de acordo com a equação (2), onde k é o número total de palavras no sistema:

$$u(c, s) = \cos(\vec{w}_c, \vec{w}_s) = \frac{\vec{w}_c \cdot \vec{w}_s}{\|\vec{w}_c\|_2 \times \|\vec{w}_s\|_2} = \frac{\sum_{i=1}^k w_{i,c} w_{i,s}}{\sqrt{\sum_{i=1}^k w_{i,c}^2} \sqrt{\sum_{i=1}^k w_{i,s}^2}} \quad (2)$$

Nesta forma, o cálculo de similaridade é realizado computando o cosseno do ângulo formado pelos dois vetores que representam os documentos (termos e frequências). Esta abordagem é utilizada no protótipo desenvolvido para calcular a similaridade entre os artigos. Por exemplo, se o usuário c , lê muitos artigos relacionados ao tópico Engenharia de Software, as técnicas de filtragem de conteúdo estão aptas a recomendar outros artigos de Engenharia de Software para o usuário c . Estes outros artigos são recomendados porque possuem termos em comum, que neste exemplo, são termos do domínio Engenharia de Software. Neste caso, o valor da função $u(c,s)$ será alto, pois os pesos calculados para os termos destes artigos (que compartilham termos em comum referente ao assunto Engenharia de Software) serão altos (Adomavicius et al., 2005).

2.1.1 Pré-Processamento de Textos

Normalmente muitas palavras de um documento não são importantes para representá-lo semanticamente. Pode-se afirmar que na maioria das vezes, substantivos ou grupo de substantivos são as palavras mais representativas em um documento. A tarefa de pré processamento, no contexto de recuperação de informação, possibilita a extração de informações úteis do texto (Cowie e Lehnert, 1996). As operações de pré-processamento abordadas e implementadas no protótipo deste trabalho são: remoção de *stopwords*, análise léxica e redução dos termos a um radical comum. Elas são detalhadas rapidamente nas seções a seguir.

2.1.1.1 Remoção de Stopwords

Stopwords são termos considerados não relevantes para a categorização ou identificação de um documento, como artigos, preposições, conjunções e pronomes. O processo de remoção de *stopwords* visa eliminar estas palavras dos textos (exemplos, “mas”, “mesmo”, “qualquer”, “seja”, “ainda”).

2.1.1.2 Análise Léxica

A análise léxica pode auxiliar a identificação de palavras no texto (Santos, 2003). Os dígitos, sinais de pontuação, acentos e hífen podem ser removidos neste processo, e todas as letras do texto podem ser transformadas para maiúsculas ou minúsculas. Neste processo de transformação das letras do texto pode ocorrer perda de informação.

2.1.1.3 Stemming

Para aprimorar as recomendações geradas pelo modelo é necessário que sejam eliminadas as variações morfológicas de cada palavra do texto de um documento (Santos, 2003). O processo de *stemming* é realizado considerando cada palavra isoladamente tentando reduzi-la a sua provável raiz. São analisadas características como grau, gênero e número com o objetivo de eliminar os sufixos e prefixos das palavras transformando-as em sua forma primitiva. Um exemplo do processo de *stemming* é o algoritmo de Porter (Porter, 2007), desenvolvidos para a língua inglesa, havendo adaptações para outras línguas. Neste trabalho utilizou-se o algoritmo de Porter (Porter,

2007) para o processo de *stemming* dos textos em língua inglesa. Para os documentos em língua portuguesa, foi utilizada uma adaptação do algoritmo de Porter, denominada de *Portuguese Stemmer* (Porter, 2007a).

3. Modelo Local

O modelo LOCAL (*LOcation and Context Aware Learning*) (Barbosa et al., 2008); utiliza informações de localização e de contexto como auxílio ao processo de ensino e de aprendizagem. No modelo LOCAL um sistema de localização acompanha a mobilidade dos aprendizes e, baseado nas suas posições físicas, explora oportunidades educacionais. O modelo MOBILREC (proposta deste trabalho e descrito na seção 4) utiliza recursos da arquitetura do LOCAL para viabilizar algumas funcionalidades e fornece serviço para o LOCAL, por isso, a importância de se descrever um pouco deste outro modelo nesta seção.

O modelo LOCAL (Barbosa et al., 2008) compõe-se basicamente de um Sistema de perfil de usuário; Sistema de localização; Assistente Pessoal; Repositório de objetos de aprendizagem; Sistema de envio de mensagens contextuais e Tutor.

4. Modelo MOBILREC

Diante da necessidade de uma *engine* de recomendação para o LOCAL, surge o modelo MobilRec que viabilizará o serviço de recomendação baseado em FBC. Os principais componentes deste modelo estão ilustrados na figura 1, localizados no quadro MobilRec do diagrama. Para seu correto funcionamento alguns agentes de software - descritos a seguir - foram desenvolvidos (Russel e Norvig, 2004): 1) O **Agente Extrator** trabalha na extração de conteúdo e na classificação de todos os artigos com extensão pdf, doc ou rtf, que estejam localizados no repositório de documentos. Todos os termos de um documento serão extraídos e inseridos em uma base de dados de termos. O processo de extração utiliza um radicalizador funcional para documentos em língua inglesa e portuguesa (Porter, 2007a). As palavras não relevantes para a análise do texto, conhecidas como *stopwords*, são desconsideradas pelo agente; 2) O **Agente Recomendador** verifica se existem novas recomendações a cada 2 horas (valor parametrizável). Para isso, o agente recomendador realiza uma chamada ao *webservice* LOCAL buscando informações relativas a usuário e perfil. De posse desta informação, o agente verifica se existem novos artigos que foram positivamente pontuados para reconfigurar os artigos que servirão de base para a geração da recomendação. Também é verificado se existem artigos inseridos após a última geração de recomendação para este usuário. As recomendações sempre são armazenadas na base de dados de recomendações, com o objetivo de fornecer um histórico e agilizar a resposta nos casos em que o agente tenha que recomendar algum artigo mesmo que o usuário não tenha feito novas pontuações em artigos e não houver novos artigos no repositório.

Se o agente recomendador identificar que alguma recomendação pode ser feita, seja por novas pontuações em artigos ou por novos artigos disponíveis no repositório, um novo acesso ao *webservice* LOCAL será realizado com o objetivo de enviar a recomendação ao sistema de mensagens do LOCAL, para que o mesmo envie a mensagem com a recomendação para o usuário.

O processo de integração entre o MobilRec e o LOCAL apresenta as seguintes etapas: 1) Um artigo científico é publicado no repositório de documentos; 2) Agente extrator identifica o novo arquivo e realiza a extração do conteúdo atualizando a base de termos e documentos, e solicita ao *webservice* LOCAL para que o mesmo insira este artigo em sua base de dados; 3) O perfil do usuário é construído através da pontuação

(nota) atribuída aos artigos; 4) O agente recomendador, que executa a cada duas horas, solicita ao LOCAL a lista de usuários e seus perfis, para verificar se existe algo a recomendar (nesta etapa, é analisado se existem novas pontuações feitas pelos usuários, com o objetivo do agente reconfigurar os artigos que servirão de base para a filtragem baseada em conteúdo, e o método TF-IDF (Garcia, 2007) é aplicado para a identificação dos artigos de maior interesse para os usuários. A geração de recomendação apenas será realizada quando existirem pontuações novas ou novos artigos no repositório de documentos); 5) O agente recomendador conecta-se no *webservice* LOCAL e envia a recomendação. O sistema de mensagens do LOCAL envia a recomendação para o usuário. O usuário ao se conectar e se identificar ao sistema usando um dispositivo móvel (*palm HP iPAQ*), receberá automaticamente todas as recomendações calculadas pelo protótipo que implementa o modelo MobilRec.

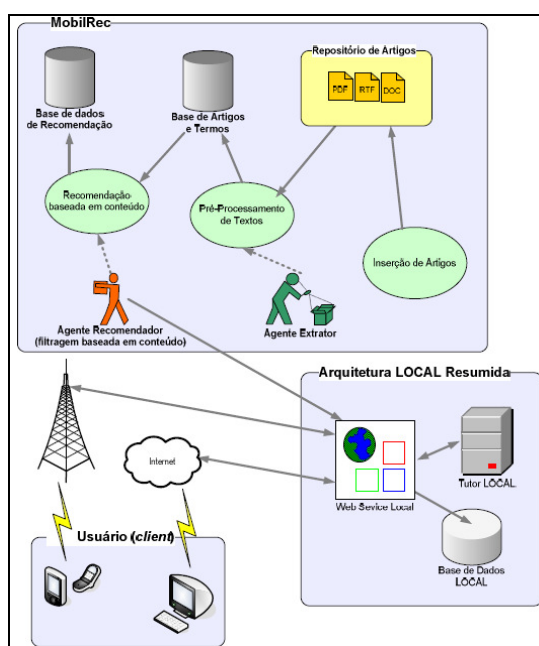


Figura 1 – Modelo MobilRec

5. Experimentos e Resultados

Objetivando avaliar o modelo desenvolvido através do protótipo implementado, foram realizados dois experimentos com alunos do curso de graduação em Engenharia da Computação da Unisinos (Universidade do Vale do Rio dos Sinos) na disciplina de Banco de Dados. Os objetivos destes experimentos envolveram verificar os valores de similaridade (artigos) calculados pelo protótipo; avaliar se as recomendações geradas pelo protótipo foram interessantes para os usuários, bem como aplicar as métricas precisão (*precision*) e abrangência (*recall*) com o objetivo de avaliar o protótipo (Silva e Cazella, 2005).

5.1 Método de Pesquisa

Os experimentos foram realizados utilizando uma amostra por conveniência (não probabilística) composta por alunos do quarto ano do curso de Engenharia da Computação da Unisinos. Desta forma foram trabalhados alunos que possuem um nível de conhecimento semelhante no que se refere a computação e banco de dados. Para iniciar a experimentação, 27 artigos foram inseridos no repositório do protótipo. Após, foram realizadas as tarefas de pré-processamento e cálculo de similaridade entre os

artigos para possibilitar a recomendação através da filtragem baseada em conteúdo. Cada estudante recebeu um e-mail no qual informava os artigos (títulos dos artigos) para serem avaliados. Este procedimento foi necessário para iniciar o sistema de recomendação, uma vez que neste nível não se dispunha de informações referentes ao gosto destes futuros usuários (problema denominado de *cold-start* em sistemas de recomendação (Cazella et al., 2010)). Uma vez que se obteve a avaliação destes alunos para os artigos selecionados, foi possível iniciar o processo de recomendação por análise de conteúdo. Posteriormente, foram utilizadas as medidas precisão e abrangência para avaliar o protótipo.

5.1.1 Métricas de avaliação

Utilizou-se as métricas Precisão e Abrangência no intuito de avaliar a capacidade do serviço de recomendação desenvolvido de retornar ao utilizador os artigos mais relevantes (*precision* ou precisão), e também a capacidade do mesmo em encontrar na coleção os artigos mais relevantes (*recall* ou abrangência). O valor máximo de precisão (1 ou 100%) é alcançado quando todos os artigos recuperados são de interesse do usuário. Por outro lado, o valor máximo de abrangência (1 ou 100%) é alcançado quando todos os artigos relevantes da coleção de artigos são recuperados. A métrica precisão é calculada da seguinte forma (Silva e Cazella, 2010):

$$\text{precisão} = \frac{\text{número de artigos relevantes recuperados}}{\text{número total de artigos recuperados}} \quad (3)$$

$$\text{abrangência} = \frac{\text{número de artigos relevantes recuperados}}{\text{número total de artigos relevantes}} \quad (4)$$

Quando se utiliza buscas por similaridade (artigos similares aos artigos avaliados positivamente pelo usuário) a tendência é aumentar significativamente o número de artigos recuperados. Para a realização dos experimentos, o grau mínimo de similaridade para a recomendação (*threshold*) foi configurado com o valor 0.1. Nos casos em que a busca recuperou artigos com similaridade ≥ 0.2 , apenas estes foram apresentados para o usuário e considerados como o número total de artigos recuperados para o cálculo da precisão. Sendo assim, quando a busca retornou artigos com similaridade ≥ 0.2 , geralmente a métrica de precisão foi melhorada, pois os artigos recuperados possuíam um grau maior de similaridade, o que pode indicar uma chance maior de acerto na recomendação. Por outro lado, a métrica de abrangência tende a piorar devido a possibilidade do protótipo não estar recomendando todos os artigos que poderiam ser interessantes para o usuário (artigos recuperados com similaridade ≥ 0.1 e ≤ 0.2).

5.2 Experimentos Realizados

5.2.1 Primeiro Experimento: Avaliação de Artigos Pré-Selecionados

Esta etapa teve como principal objetivo permitir que os usuários avaliassem os artigos que lhes foram designados através do uso de uma escala *likert* de 5 pontos (1- Péssimo e 5 – Excelente). Estas avaliações serviram de base para a geração das recomendações do segundo experimento. Cada estudante recebeu um email no qual foi informado o título de 5 artigos para serem avaliados. Estes artigos foram selecionados de forma aleatória. Os estudantes foram cadastrados no protótipo e receberam suas credenciais (usuário e senha) para efetuarem o *login* na interface e iniciarem as avaliações. Das 50 avaliações propostas (5 avaliações por usuário) foram realizadas 46, o que equivale a 92% de avaliações realizadas.

5.2.2 Segundo Experimento: Geração das Recomendações

Com base nas avaliações realizadas pelos usuários no primeiro experimento, o protótipo gerou as recomendações para os usuários. Neste experimento, os usuários visualizaram as recomendações geradas, para posteriormente avaliarem essas recomendações. As avaliações destas recomendações foram realizadas da mesma maneira que a avaliação do primeiro experimento, ou seja, os usuários avaliaram os artigos que receberam como recomendação de acordo com a escala *likert* de 5 pontos. O protótipo recomendou artigos com similaridade ≥ 0.2 aos artigos avaliados positivamente pelos usuários. Nos casos em que nenhuma recomendação foi gerada com o valor mínimo de similaridade 0.2, o valor foi alterado para 0.1 para possibilitar alguma recomendação. Na tabela 1, são apresentados os valores de similaridade entre alguns dos artigos.

Tabela 1 - Artigos similares

Artigo 1	Artigo 2	Similaridade
A2.pdf	A3.pdf	0.2282
A5.pdf	A8.pdf	0.1068
A10.pdf	A12.pdf	0.1066
A10.pdf	A24.pdf	0.3174
A10.pdf	A27.pdf	0.5212
A11.pdf	A8.pdf	0.2185
A12.pdf	A14.pdf	0.1206
A15.pdf	A8.pdf	0.1059
A15.pdf	A23.pdf	0.1173
A15.pdf	A24.pdf	0.1039
A16.pdf	A17.pdf	0.1157

Tabela 2- Recomendações geradas

Artigos / Usuários	U1 (0.2)	U2 (0.2)	U3 (0.2)	U4 (0.1)	U5 (0.2)	U6 (0.1)	U7 (0.2)	U8 (0.2)	U9 (0.1)	U10 (0.1)
A1					5				3	
A4				2						
A5		4			5	5				4
A6	3		1							
A7			1	2						
A8	R+	R-	R-	R+	R+	R+	R-		3	2
A9						4				
A10		R+						5		
A11	4				5					
A12		3				3	R-	R-	R+	
A13				4						
A14							4		4	
A15		R-		4						
A16	4	R+	4			R+	4		R+	
A17	R-	R-	3			R+	R-		R+	

De uma maneira geral, os valores de similaridade entre os artigos da coleção se mostraram baixos (média do grau de similaridade é 0,0582), considerando que o grau de similaridade pode variar entre [0;1]. Isto explica a utilização do valor mínimo 0.1 ou 0.2 (dependendo do usuário) para a geração das recomendações, ou seja, utilizou-se um grau mínimo de similaridade baixo para possibilitar que o protótipo recomende os artigos para os usuários. As recomendações geradas pelo protótipo, resultado deste experimento, podem ser vistas na tabela 2. Elas estão representadas através das descrições 'R+' e 'R-' nas células da tabela. As células da tabela com a descrição 'R+' indicam as recomendações que foram enviadas aos usuários, ou seja, que possuem valor de similaridade calculado maior ou igual ao grau mínimo definido (os valores entre parênteses, nas células referentes aos usuários, indicam o grau mínimo de similaridade para o protótipo gerar as recomendações), enquanto que as células com a descrição 'R-' indicam uma recomendação encontrada, porém não enviada ao usuário devido à similaridade calculada ser menor que o grau mínimo definido. O protótipo enviou um total de 34 recomendações para os usuários (células com descrição 'R+'), as demais células com valores numéricos referem-se a avaliações feitas pelos usuários aos itens.

As avaliações realizadas pelos usuários para estas recomendações que foram enviadas pelo protótipo (representadas pela descrição 'R+' nas células da tabela 2) estão parcialmente ilustradas na tabela 3, através da descrição 'R+' acrescida da pontuação que o usuário informou. As descrições 'R+' que não estão acrescidas de alguma pontuação indicam que o usuário não avaliou a recomendação recebida. As recomendações não avaliadas foram desconsideradas para o cálculo de precisão e abrangência. Para o cálculo das métricas deste experimento, deve-se observar o seguinte: 1) Foi considerado como o número total de artigos relevantes o total de recomendações enviadas que foram avaliadas pelos usuários (células com a descrição 'R+' acrescida da pontuação) somado ao número de recomendações encontradas porém

não enviadas (células com a descrição ‘R-’); 2) O número total de artigos recuperados equivale ao número total das recomendações enviadas que foram avaliadas pelos usuários (células com a descrição ‘R+’ acrescida da pontuação); 3) O número total de artigos relevantes recuperados é o número de recomendações (artigos) avaliadas positivamente (4 – Bom e 5 – Excelente). Das 34 recomendações enviadas, foram avaliadas 25, ou seja, 73,52% das recomendações enviadas foram avaliadas pelos usuários. Os valores de precisão e abrangência (tabela 4) apresentam uma melhora sobre os valores calculados no primeiro experimento (tabela 3). A média de precisão do primeiro experimento foi 0.62, enquanto que no segundo experimento foi de 0.77. A média de abrangência do primeiro experimento (tabela 3) foi de 0.11, enquanto que no segundo experimento (tabela 4) foi de 0.49. Isso mostra que além de melhorar a qualidade da recomendação, o protótipo identificou um maior número de artigos relevantes para os usuários na coleção de artigos. Pode-se observar que para os usuários *U6* e *U10* (tabela 4) obteve-se precisão e abrangência de 100%, ou seja, todos os artigos relevantes da coleção para estes usuários foram encontrados (recomendados) e todas as recomendações realizadas foram relevantes para eles. Já para o usuário *U3*, as duas medidas ficaram com o valor 0. O valor 0.2 para o grau mínimo de similaridade fez com que o protótipo encontrasse apenas 1 recomendação e esta foi avaliada como regular pelo usuário, que para este protótipo, indica que a recomendação não é relevante. Isso explica o valor 0 para as duas medidas. Desta maneira, pode-se observar que ao tentar aumentar a precisão da recomendação, ou seja, aumentar o grau mínimo de similaridade no intuito de recuperar somente artigos que o usuário tenha interesse, a abrangência tende a piorar, pois o protótipo pode deixar de enviar recomendações que poderiam ser relevantes.

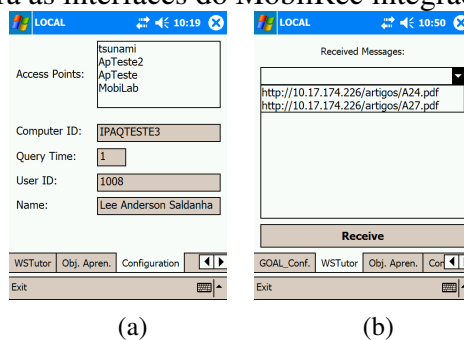
Tabela 3 - Avaliação Experimento 1

Artigos / Usuários	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
A15				4				X		
A16	4		4				4			
A17			3							
A18								X		4
A20						4		X	4	
A21	5				4		4			
A22					3					5
A23		5								
A24										
A25										4
A26				4			4			
A27		4					3			
A28		3	4							
A29					2					
A30	5									4
Precisão	0,8	0,6	0,4	0,6	0,8	0,6	0,8	0,2	0,6	0,8
Precisão Média	0,62									
Abrangência	0,14	0,11	0,07	0,11	0,14	0,11	0,14	0,03	0,11	0,14
Abrangência Média	0,11									

Tabela 4 - Avaliação Experimento 2

Artigos / Usuários	U1 (0,2)	U2 (0,2)	U3 (0,2)	U4 (0,1)	U5 (0,2)	U6 (0,1)	U7 (0,2)	U8 (0,2)	U9 (0,1)	U10 (0,1)
A15		R-		4						
A16	4	R+ 5	4			R+	4		R+ 4	
A17	R-	R-	3			R+	R-		R+	
A18										4
A20	R-	R-	R-			4	R-		4	
A21	5	R-	R-		4		4		R+	
A22	R-		R-	R+	R-	3	R-		R+ 4	5
A23	R+ 3	5	R+ 3	R+ 4		R+	R+ 4		R+ 4	
A24	R-	R+		R+ 4	R-	R+ 4	R-	R+ 5		
A25	R-		R-	R+	R-		R-		4	R+ 5
A26			R-	4			4		R+ 5	R+ 5
A27	R-	4	R-		R-		3	R+ 4		
A28	R+ 5	3	4	R+ 3	R-		R+ 3		R+ 3	R+ 4
A29					2					
A30	5									4
Precisão	0,66	1	0	0,75	1	1	0,5	1	0,8	1
Precisão Média	0,77									
Abrangência	0,22	0,28	0	0,75	0,16	1	0,11	0,66	0,8	1
Abrangência Média	0,49									

A figura 2 ilustra as interfaces do MobilRec integrado com o LOCAL.



(a)

(b)

Figura 2 – (a) Interface de login (b) Interface acesso a recomendação do MobilRec

6. Conclusão

Neste trabalho, foi apresentado um protótipo de um modelo que viabiliza o serviço de recomendação baseado em filtragem de conteúdo para aprimorar as recomendações de artigos realizadas pelo framework LOCAL. O protótipo foi desenvolvido através de um agente de pré-processamento de textos, um agente de recomendação e um *webservice client* para a integração com o serviço de mensagens do LOCAL. Os resultados dos experimentos realizados com o protótipo mostraram que a filtragem baseada em conteúdo pode ser muito eficiente para a recomendação de itens textuais. Com base na primeira avaliação de alguns artigos realizada pelos usuários, o protótipo recomendou novos artigos que foram avaliados em sua maioria, como interessantes pelos usuários. O grau mínimo de similaridade para a recomendação foi configurado com o valor 0.2 ou 0.1, para permitir a geração das recomendações, considerando que a amostra selecionada poderia conter artigos com similaridades próximas de zero. Os valores das medidas precisão e abrangência, apresentados, auxiliaram a comprovar a eficiência da solução proposta. É importante ressaltar que os resultados obtidos se aplicam a amostra não-probabilística utilizada, não é possível afirmar que a solução apresentará o mesmo resultado para outra amostra.

Alguns possíveis trabalhos futuros compreendem: 1) Avaliar e comparar os resultados obtidos utilizando outras medidas de cálculo de similaridade entre artigos científicos; 2) Considerar apenas os termos mais representativos de um artigo na aplicação do modelo de espaço vetorial e cálculo de similaridade através da medida do cosseno, com o objetivo de melhorar o grau de similaridade calculado para documentos extensos; 3) Avaliar a utilização de outros algoritmos de *stemmer* para o processo de radicalização dos termos dos documentos; 4) Aprimorar a análise léxica dos textos. Na tarefa de pré-processamento do texto, mais especificamente na extração do conteúdo dos artigos, pode ser criado um tratamento mais abrangente para caracteres especiais.

7. Referências

- Adomavicius, G., Tuzhilin A., Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions, *IEEE Transactions on Knowledge and Data Engineering*, 17, 6 (Jun. 2005), 734-749.
- Balabanovic, M., Shoham, Y. Fab: Content-Based, Collaborative Recommendation. *Communications of the ACM*, 40, 3 (Mar. 1997), 66-72.
- Barbosa, J. L. V., Hahn, R.,; Rabello, S. A, Barbosa, D. N. F. . LOCAL: a Model Geared Towards Ubiquitous Learning. In: *39th ACM Technical Symposium on Computer Science Education (SIGCSE)*, 2008, Portland. Proceedings of the ACM SIGCSE 2008. New York : ACM Press, 2008. 432-436.
- Cazella, S. C. ; Nunes, M. A. ; Reategui, Eliseo . A Ciência da Opinião: Estado da arte em Sistemas de Recomendação. In: André Ponce de Leon F. de Carvalho; Tomasz Kowaltowski. (Org.).. (Org.). JAI: *Jornada de Atualização em Informática da SBC*. Rio de Janeiro: Editora da PUC Rio, 2010, v. , p. 161-216.
- Cowie, J., W. Lehnert, Information Extraction. *Communications of the ACM*, vol. 39, New York, USA, Jan. 1996. p. 80-91.
- Garcia, E. **The Classic Vector Space Model**: Description, Advantages and limitations of the Classic Vector Space Model. [S. l.: s.n.], 2005. Disponível em: <<http://www.miislita.com/term-vector/term-vector-3.html>>. Acessado em: 10 de jul. de 2007.
- Herlocker, J., Konstan, J., Riedl, J., Explaining Collaborative Filtering Recommendations. In: *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, 241-250.
- Porter, M. *An algorithm for suffix stripping*. Originally published in Program, 14 no. 3, p. 130-137, July 1980. Disponível em: <<http://www.tartarus.org/~martin/PorterStemmer/def.txt>>. Acessado em: 05 de abr. de 2007.



Porter, M. *Portuguese stemming algorithm*: The stemming algorithm. 2005. Disponível em: <<http://snowball.tartarus.org/index.php>>. Acessado em: 05 de abr. de 2007a.

Russell, Stuart J.; Norvig, Peter. *Inteligência artificial*, Rio de Janeiro: Elsevier, 2004. 1021 p.

Salton, G., McGill J. M. *Introduction to Modern Information Retrieval*, McGraw Hill, New York, 1983.

Santos, M. A. M. R. *Extraindo regras de associação a partir de textos*. 2002. 51 f. Dissertação (Mestrado em Informática Aplicada) - Programa de Pós-Graduação em Informática Aplicada, Pontifícia Universidade Católica do Paraná.

Schafer, J. B. et al., Recommender system in Ecommerce, In: *Proceedings of the ACM Conference on Electronic Commerce* (EC' 99), N. Y., USA, 158-166.

Silva F. W., Cazella, S. C., STAR: Um Framework para recomendação de artigos científicos baseado na relevância da opinião dos usuários e em filtragem colaborativa. In: *Encontro Nacional de Inteligência Artificial* (ENIA), 2005, São Leopoldo, RG, Brazil.