Tecnologias de SW livre para análise de mensagens em chats e recomendações online de documentos eletrônicos

Rodrigo Branco Kickhöfel[1], Stanley Loh[1,2], Daniel Lichtnow[1], Ramiro Saldaña[1], Thyago Borges[1], Tiago Primo[1], Gabriel Simões[1], Gustavo Piltcher[1]

[1]Universidade Católica de Pelotas (UCPEL) – Escola de Informática
Rua Felix da Cunha, 412 – Pelotas – RS – Brasil

[2]Universidade Luterana do Brasil (ULBRA) – Faculdade de Informática
R. Miguel Tostes, 101, Canoas, RS – Brasil

{rodrigok,lichtnow,rsaldana,thyago,gustavopil}@ucpel.tche.br,
sloh@terra.com.br, tiagoprimo@brturbo.com.br, gsimoes@vetorial.net

Abstract: *This paper describes a recommender system to support collaboration using open source technologies. The system consists on a WebChat that using text mining techniques identifies the subject of the messages and selects items from a digital library to recommend to the chat participants. The system allows knowledge exchange among members of virtual communities. The results and the advantages of using open source technologies for that purpose are presented and discussed.*

Resumo. *Este artigo apresenta um sistema de recomendação para apoio à colaboração utilizando tecnologias de software livre. O sistema consiste em um chat Web que, a partir de técnicas de text mining, identifica o assunto das mensagens e gera recomendações a partir de uma base de dados própria (uma Biblioteca Digital, contendo documentos eletrônicos, links par páginas web e referências bibliográficas). O sistema permite a troca de conhecimento entre os membros de comunidades virtuais. São apresentados os resultados obtidos e as vantagens da utilização da tecnologia de software livre para estes propósitos.*

1. Introdução

A geração de novos conhecimentos é desencadeada em parte pela troca de conhecimento entre as pessoas, por exemplo, através de *chats* na *Web*. Então com o objetivo de valorizar a troca do conhecimento, surgiram os sistemas de recomendação (*Recommender Systems*). Um sistema de recomendação tem por finalidade auxiliar no processo social de indicar ou receber indicação, seja esta indicação referente a livros, artigos, discos, restaurantes, ou informações (RESNICK & VARIAN, 1997).

Este artigo discute o uso de tecnologias de software livre para auxiliar ou suportar um sistema de recomendação para apoio à colaboração em comunidades virtuais.

2. O Sistema SisRecCol

O SisRecCol (Sistema de Recomendação para apoio a Colaboração) consiste em um *chat Web* onde os usuários trocam mensagens e recebem recomendações conforme o assunto identificado nas mensagens. Essa identificação é realizada através de técnicas de *text mining* apoiadas em uma ontologia desenvolvida para o sistema. As recomendações são de documentos eletrônicos e *links* para *sites Web* (armazenados numa Biblioteca Digital privada do sistema), de autoridades nos assuntos (identificados pelo próprio sistema) e de discussões anteriores sobre o mesmo tema. O objetivo é apoiar a troca de conhecimento entre os usuários. Dessa forma a ferramenta pode ser utilizada em

grupos que pertençam a um domínio específico, como turmas de ensino a distância, auxiliando inclusive na avaliação dos alunos, permitindo analisar a interação de cada um e seus assuntos de interesse. O sistema pode ser adaptado para ser utilizado em diferentes áreas, bastando incluir uma nova ontologia, referente ao novo domínio. O protótipo do sistema está disponível em http://gpsi.ucpel.tche.br/sisrec.

2.1 Funcionamento

O sistema utiliza uma ontologia de domínio para classificar os documentos na biblioteca digital, para traçar o perfil dos usuários e para identificar temas nas mensagens (através de técnicas de *text mining*). Uma ontologia do domínio (*domain ontology)* é uma descrição de "coisas" que existem ou podem existir em um domínio (SOWA, 2002) e descreve o vocabulário relacionado ao domínio em questão (GUARINO, 1998).

No SisRecCol a ontologia é apresentada como uma hierarquia de conceitos e cada conceito contém uma lista de termos associada. No momento, há somente uma ontologia no Sistema, pertinente à área da Ciência da Computação e contendo termos em Inglês e Português, o que permite discussões em ambas as línguas.

O módulo de *text mining* é responsável pela identificação dos assuntos das mensagens. Essa identificação não ocorre apenas individualmente nas mensagens, mas sim analisando um contexto (se várias mensagens tratam do mesmo assunto e com um grau de importância acima de um determinado parâmetro, a probabilidade desse assunto ser o tópico da discussão é alta, sendo esse então identificado como assunto principal desse conjunto de mensagens).

Biblioteca Digital é uma coleção de recursos digitais, organizados sob uma certa lógica, acessíveis para recuperação sobre uma rede de computadores (KOCHTANEK; HEIN e KASSIM, 2001). A Biblioteca Digital do SisRecCol contém documentos eletrônicos e *links* para páginas *Web*, no momento. Este conteúdo é previamente indexado com ferramentas de software que realizam o processo de forma automática com base nos mesmos métodos usados para identificar os assuntos das mensagens do *chat*. O processo de indexação determina o grau de relacionamento dos documentos, (artigo ou *site*) aos conceitos existentes na ontologia.

A inclusão dos documentos na Biblioteca Digital pode ser feita pelos usuários, num processo independente da sessão do *chat*. O material disponibilizado pelos usuários passa pela revisão do administrador do sistema, que verifica a qualidade do material e também avalia a indexação produzida pelo sistema. Uma vez aprovado, o documento passa a estar disponível para utilização, podendo ser recomendado durante as sessões do c*hat*, ou também ser consultado utilizando palavras-chaves ou a estrutura hierárquica da ontologia.

A base de perfis do sistema se assemelha aos chamados Mapas do Conhecimento ou Páginas Amarelas (*Yellow Pages*), que servem para indicar que pessoas possuem determinados conhecimentos. Ela contém dados de identificação dos usuários e suas áreas de conhecimento, representadas pelos conceitos presentes na ontologia. A definição do perfil é dinâmica, ou seja, conforme o usuário utiliza o sistema, ele é pontuado em determinados assuntos (conceitos) conforme suas ações. A base de perfis é utilizada na definição dos itens que serão recomendados para cada usuário, e também é útil na avaliação dos membros do grupo.

A partir do assunto identificado nas mensagens do *chat,* o módulo de recomendação proativamente sugere uma lista de conteúdos relevantes, vindos da Biblioteca Digital (documentos eletrônicos, *sites Web* ou referências bibliográficas) aos usuários, de acordo com o perfil de cada um (*content-based recommendation*) ou avaliando as ações das autoridades no sistema (*collaborative filtering*). Essa recomendação é apresentada em uma área separada, não interferindo no decorrer da discussão, e permitindo ao usuário acessar o conteúdo (por exemplo, ler um documento, abrir um *site Web* ou ver dados das autoridades no conceito identificado).

2.2 Tecnologias Utilizadas

O custo de aquisição de software e a facilidade para encontrar documentação detalhada, principalmente do sistema operacional do servidor e o banco de dados é um importante fator a favor do software livre. Optou-se por utilizar a plataforma Linux para o desenvolvimento do sistema. Como ele consiste em um *chat*, que precisa ser acessado via *Web*, a escolha mais natural é o Apache como servidor *Web*, por estar diretamente associado ao Linux, além de ser um dos mais difundidos e de fácil utilização.

O SisRecCol utiliza um banco de dados para armazenar a ontologia, as informações da Biblioteca Digital, os perfis de usuário bem como todas as informações referentes as sessões do *chat*. Em função da adoção do Linux, duas opções principais de banco de dados livre atenderiam as necessidades do projeto, MySQL ou PostgreSQL. Apesar do intenso uso do banco de dados, não se trata de uma aplicação crítica. A velocidade é um fator importante e por isso se optou pelo MySQL, que, apesar de ter menos recursos que o PostgreSQL, é um banco bem ágil, muito difundido e atendeu perfeitamente as necessidades do sistema. Utilizando tabelas do tipo MyISAM, o desempenho do MySQL favoreceu o suporte a um grande número de conexões simultâneas, fundamental para a proposta do sistema. Como este tipo de tabela não implementa integridade referencial ou controle de transações, as respostas do banco ficaram extremamente rápidas. Só para se ter uma idéia, a cada mensagem enviada por um participante do *chat*, o sistema precisa:

1) comparar esta mensagem com todos os conceitos presentes na ontologia (lembrando que cada conceito é representado por uma lista de palavras e pesos associados), para identificar o assunto;

2) avaliar se o assunto é válido, analisando o contexto (um grupo de mensagens mais recentes);

3) recuperar da Biblioteca Digital, os itens classificados a este assunto identificado;

4) avaliar, para cada usuário, se os itens serão recomendados ou não, já que o usuário pode já ter "baixado" o documento ou "acessado" o *site*

5) então apresentar para cada usuário individualmente a lista de recomendações

6) além disto, o usuário pode simultaneamente abrir ou "baixar" os documentos recomendados

7) e ainda cada usuário recebe pontos por participar do *chat*, por enviar mensagens, por baixar documentos, etc.

A linguagem de programação utilizada foi o PHP. Ela se mostrou a melhor opção no momento, pois em termos gerais (e se tratando de aplicações *Web*) tem um bom desempenho comparado a outras linguagens como Java, e é de fácil utilização tendo uma boa documentação. Outra vantagem é que funciona em várias plataformas, tornado o sistema mais portável. Como o sistema trabalha com processamento de textos, o desempenho é uma questão importante ainda mais se tratando de um *chat* em tempo real. Nos testes realizados com até 10 usuários o PHP apresentou um bom resultado de resposta, gerenciando bem as diversas conexões e não prejudicando a discussão no c*hat*. Além disso, PHP possui suporte nativo ao banco MySQL, como já citado, o SGBD utilizado pelo SisRecCol.

O ambiente de *chat* necessita de algumas validações, além disso, precisa enviar e mostrar vários dados sem que o conteúdo desapareça da tela do usuário no momento da submissão. Sendo assim, o SisRecCol utiliza uma série de *scripts* em JavaScript para validar, enviar e receber dados do servidor. Essa tecnologia foi amplamente utilizada já que a maioria dos *browsers* atuais possui suporte nativo, eximindo o usuário da tarefa de instalar qualquer API ou *virtual machine* para utilizar o sistema.

3. Experimentos

Foram realizados experimentos de discussões com grupos utilizando o sistema proposto. As pessoas do grupo foram convidadas a entrar no *chat* num horário determinado e realizar uma reunião

online real. Em uma das sessões, que durou aproximadamente 1 hora, o sistema registrou 374 mensagens (uma média de 6,23 mensagens por minuto, e quase uma mensagem a cada 10 segundos). Uma avaliação formal do método de identificação de assuntos registrou uma precisão de 85,7% (ou seja, porcentagem de assuntos corretamente identificados) e uma abrangência de 85,7% (ou seja, porcentagem de assuntos que deveriam ter sido identificados), o que leva a crer que, além de rápido, o sistema consegue um bom nível de acerto na análise das mensagens.

Até então, nos testes realizados com até 10 usuarios, o ambiente de *chat*, assim como o módulo de recomendação mostraram-se totalmente estáveis, apresentando em alguns momentos *delays*, referentes a instabilidades na rota de rede utilizada pelo cliente. Uma biblioteca digital com um número maior de documentos ou um volume de usuários maior pode causar alguma perda de desempenho, visto que a recomendação é complexa (as mensagens são analisadas em função da ontologia, e, além disso, é preciso analisar o perfil dos usuários, os itens que serão recomendados para cada usuário e então gerar a recomendação nas telas desses usuários).

Notaram-se também algumas incompatibilidades com *browsers* mais antigos, os quais não suportam algumas implementações DOM (*Document Object Model*) utilizadas nas atuais versões de JavaScript.

4. Conclusão

As tecnologias de software livre se mostraram totalmente adequadas ao propósito do sistema desenvolvido e por isso a escolha foi correta. Os resultados obtidos foram muito bons tanto em rapidez quanto em consistência e confiabilidade. Uma possível alteração para o futuro seria a utilização da linguagem Java para criação de novas funcionalidades, devido a sua grande quantidade de recursos, como por exemplo, *multithreading* para o gerenciamento das várias conexões. A troca do banco de dados talvez seja necessária conforme aumente a quantidade de informações armazenadas.

Vale ressaltar que as tecnologias baseadas em software livre utilizadas no SisRecCol adaptam-se completamente também a outros ambientes operacionais, alguns até mesmo proprietários, sem significativa perda de performance ou necessidade de instalação de pacotes ou *frameworks* adicionais, favorecendo assim, a disponibilidade do sistema pare uma gama muito grande de usuários, espalhados por todos os nodos da rede.

Agradecimentos

O presente trabalho foi realizado com o apoio do CNPq, uma entidade do Governo Brasileiro voltada ao desenvolvimento científico e tecnológico.

Referências Bibliográficas

GUARINO, Nicola (1998) Formal Ontology and Information Systems. In: International Conference on Formal Ontologies in Information Systems - FOIS'98, Trento, Itália, Junho de 1998. p.3-15

KOCHTANEK T.R.; HEIN K.K., KASSIM A. R. C. (2001) “A digital library resource Web site: Project DL”, Online Information Review, v.25, n.1.

RESNICK, P. & VARIAN, H. (1997) Recommender systems. Communications of the ACM, v.40, n.3, Março, p.56-58.

SOWA, J. F. (2002) “Building, sharing, and merging ontologies”, AAAI Press / MIT press, p.3-41.