Análise Automática de Temas Discutidos por *Chat* em uma Comunidade Virtual

Daniel Lichtnow[1], Stanley Loh[1,3], Luiz Antônio M. Palazzo[2], Ramiro Saldaña[1], Thyago Borges[1], Tiago Primo[1], Rodrigo Branco Kickhöfel[1], Gabriel Simões[1]

[1]Universidade Católica de Pelotas (UCPEL) – Grupo de Pesquisa em Sistemas de Informação R. Félix da Cunha, 412, Pelotas, RS – CEP 96010-000

[2]Universidade Católica de Pelotas (UCPEL) – Grupo de Pesquisa em Inteligência Artificial R. Félix da Cunha, 412, Pelotas, RS – CEP 96010-000

[3]Universidade Luterana do Brasil (ULBRA) – Faculdade de Informática R. Miguel Tostes, 101, Canoas, RS – CEP 92420-280

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

Resumo. *A Internet possibilita a criação de Comunidades Virtuais, para compartilhamento de conhecimento entre pessoas distantes fisicamente. Embora existam várias tecnologias para apoiar estas comunidades, o fato é que o conhecimento compartilhado nestas comunidades poderia ser melhor aproveitado. Dentre estas possibilidades está a análise dos assuntos discutidos por membros de uma Comunidade Virtual. Este artigo descreve um sistema que identifica os assuntos abordados em um chat Web privativo e a partir disto, realiza a recomendação de conteúdos. Posteriormente, a análise estatística dos temas discutidos permite descobrir conhecimento sobre a comunidade e seus membros.*

Abstract. *Internet enables the raising of Virtual Communities, where people can share knowledge without physical frontiers. Despite all the technologies available to support these communities, the knowledge shared among theirs members could be better utilized. One of these possibilities is to analyze the themes present in discussions of a community in order to evaluate members, trends of the group, depth and coverage of the discussion. This paper describes a system that identifies themes discussed in a private Web chat and then recommends items from a digital library during the discussion. After the discussion session, statistical analyses allows discovering knowledge about the community and its members.*

1. Introdução

A troca de conhecimento entre os membros de uma organização contribui para a produção de inovações e para aprimorar o conhecimento de cada um dos seus membros [Nonaka e Takeuchi 1995]. Esta percepção de que as organizações e os grupos de usuários podem se beneficiar de uma base de conhecimento compartilhada serviu como incentivo para criação das Comunidades Virtuais.

Uma Comunidade Virtual corresponde a um grupo de indivíduos (os membros da comunidade ou usuários) que dividem conhecimento, interesses e objetivos em um domínio específico através da Internet. Exemplos de comunidades virtuais são as constituídas pelos membros de quaisquer organizações em espaços *online*, *intranets* corporativas, ambientes de trabalho colaborativo, sistemas educacionais *online*, etc [Palazzo e Costa 2000].

Estes membros devem participar espontaneamente, com objetivos comuns através de relações e interações sociais e utilizando um repertório comum [Agostini et al. 2003]. Em uma Comunidade Virtual procura-se oferecer aos membros conteúdos e ferramentas que permitam a comunicação síncrona ou assíncrona. Através desta combinação entre conteúdo e ferramentas, procura-se fazer com que os membros de uma Comunidade Virtual adquiram novos conhecimentos relacionados as suas áreas de interesse.

Considerando a situação do Brasil, onde existe concentração de profissionais e de conhecimento em algumas regiões, a criação de Comunidades Virtuais se apresenta como uma possibilidade de minimizar estas diferenças mediante a troca de conhecimento e acesso facilitado a conteúdo digital.

Dentre as fontes de conhecimento existentes dentro de uma comunidade virtual estão, por exemplo, os *sites*, os artigos, os livros, e os membros com determinado grau de conhecimento no assunto.

Parte deste conteúdo pode ser produzido pelos próprios membros da comunidade sendo que em muitos casos o conhecimento pertencente a alguns membros não é encontrado de forma estruturada, não está codificado ou não é adequadamente reutilizado.

Pode-se citar como exemplo de conhecimento que não é freqüentemente bem utilizado, aquele conhecimento produzido ou compartilhado pelos membros de uma comunidade virtual a partir das mensagens colocadas em um *chat online*. Geralmente, este conhecimento só é retido pelos membros individualmente e não é analisado para o benefício do grupo todo. Este conhecimento trocado pelos membros através de uma *chat* poderia ser melhor aproveitado se ocorresse a análise do conteúdo das mensagens. A identificação dos assuntos de interesse de um usuário permitiria aprimorar a adaptação do conteúdo existente na Comunidade Virtual aos interesses de seus membros, servindo também como importante apoio para aprimorar aspectos ligados a *filtragem social*, processo no qual, os interesses são determinados indiretamente pela distribuição dos usuários segundo grupos de interesses comuns [Maes 1994] [Resnick e Varian 1997]. Além disto, seria possível descobrir os membros da comunidade com determinados níveis de conhecimento em certos assuntos.

Partindo da idéia de fornecer ferramentas que facilitem as atividades de cooperação e colaboração e que também facilitem a disseminação do conhecimento entre membros de uma comunidade virtual que estejam dispersos (residindo em diversas regiões do Brasil, por exemplo), este artigo apresenta a proposta de um sistema que consiste em um *chat* que além de permitir a troca de mensagens entre seus usuários, procura identificar os assuntos das mensagens e além disto fazer recomendações de conteúdos (artigos, *sites*, discussões anteriores e usuários com determinados conhecimentos) relacionados aos assuntos identificados a serem consultados pelos usuários durante a sessão do *chat* ou posteriormente quando lhes for mais oportuno.

Espera-se mediante a recomendação destes conteúdos oferecer aos usuários alguns subsídios úteis à discussão assim como facilitar a compreensão e o aprofundamento dos assuntos tratados em uma sessão de *chat*. Além disto, espera-se aproveitar melhor o conteúdo existente no repositório de uma comunidade virtual assim como definir com mais precisão o perfil e os interesses de cada membro da comunidade.

A seção 2 deste artigo apresenta a descrição do sistema implementado, sendo destacados cada um dos seus módulos. A seção 3 relaciona o sistema a outras propostas e sistemas implementados. Já na seção 4 é feita uma avaliação do sistema, sendo esta relacionada sobretudo a capacidade do sistema de identificar assuntos e de recomendar conteúdo útil. Na seção 5 são descritos experimentos feitos com o sistema, que demonstram algumas das possibilidades que o sistema oferece no que se refere ao entendimento da interação dos usuários. Finalmente, na seção 6 são apresentadas as conclusões.

2. Descrição do Sistema

O objetivo principal do sistema é identificar os assuntos tratados em um *chat* para a partir desta identificação realizar a recomendação de conteúdos relevantes a estes usuários.

Uma vez que o sistema irá identificar os assuntos tratados nas sessões de *chat*, será possível definir com maior precisão quais são os interesses dos usuários.

O módulo de *Text Mining* identifica os assuntos tratados a partir da análise dos termos presentes nas mensagens e da comparação destes com uma ontologia relacionada aos assuntos (conceitos) de interesse dos usuários.

Feita a identificação do assunto, este é passado para o módulo de recomendação que irá selecionar itens existentes na Biblioteca Digital, classificados no mesmo assunto. Ao fazer uma recomendação, o sistema irá verificar o perfil de cada usuário participante da sessão, tentando determinar se os conteúdos relacionados ao assunto deverão ou não ser recomendados a ele.

É importante salientar, que não é necessário que um usuário tenha o seu perfil previamente definido para começar a utilizar o sistema. A existência do perfil irá servir para aprimorar o processo de recomendação.

A partir da classificação proposta em [Terveen e Hill 2001], o sistema aqui proposto pode ser classificado como um sistema de recomendação baseado em conteúdo - *content-based recommender*, isto porque os assuntos identificados nas mensagens irão gerar recomendações de conteúdos relacionados a estes assuntos.

A versão atual do sistema foi implementada utilizando as tecnologias livres como linguagens de programação PHP e Javascript, banco de dados MySql, servidor web Apache e sistema operacional Linux. Um protótipo do sistema encontra-se disponível no endereço http://gpsi.ucpel.tche.br/sisrec. É preciso fazer um cadastro inicial para utilizar o sistema, para que possa ser identificado o perfil do usuário. Os módulos serão descritos em detalhe nas próximas seções.

2.1 O *chat*

O *chat* funciona como os diversos *chats* existentes, mas exige o registro de cada usuário para que possa ser utilizado.

A principal limitação da versão atual reside no fato de que não é oferecida a possibilidade de que sejam usados vários canais de discussão. Este recurso estará presente em versões posteriores.

2.2 O módulo de *Text Mining*

O módulo de *Text Mining* trabalha como um *sniffer,* examinando cada uma das mensagens enviadas pelos usuários. Os assuntos ou temas são identificados pela comparação de alguns termos que aparecem nas mensagens, com termos que estão relacionados a conceitos presentes na ontologia do sistema.

O método utilizado neste sistema (que é um tipo de classificação) foi apresentado em [Loh; Wives e Oliveira 2000]. Este método não utiliza técnicas de Processamento de Linguagem Natural para analisar a sintaxe e a semântica de cada mensagem, sendo um método que está baseado em técnicas probabilísticas

Assim, considerando uma Comunidade Virtual formada por usuários da área de computação, o termo “neural” dentro de uma mensagem indicaria que o assunto provável da mensagem está relacionado ao assunto/tema (conceito) Redes Neurais e Inteligência Artificial.

O algoritmo usado é baseado nos algoritmos de Rocchio e Bayes [Rocchio 1996] [Ragas e Koster 1998] [Lewis 1998] e utiliza vetores para representar textos e conceitos. Os vetores que representam os textos e os conceitos são compostos por uma coleção de termos com um peso associado a cada termo. No caso dos textos que compõe as mensagens, o peso de cada termo é dado pelo cálculo da freqüência relativa de cada termo no texto, isto é, o número de ocorrências de um termo dentro de uma mensagem dividido pelo número total de termos da mensagem. Já o peso de um termo em relação a um conceito representa a probabilidade que o termo tem de indicar um determinado assunto. Na montagem dos vetores são ignoradas as chamadas *stopwords* - termos que aparecem com muita freqüência como as preposições, por exemplo, - e que não tem relevância na identificação dos assuntos de uma discussão. Detalhes sobre como este peso é obtido são descritos na seção sobre ontologia.

O método avalia a similaridade entre um texto e um conceito presente na ontologia usando uma função de similaridade que calcula a distância entre dois vetores.

A função de similaridade multiplica os pesos dos termos que estão presentes nos dois vetores, sendo que a soma destes produtos, limitada a 1, é o grau de similaridade existente entre o texto e o conceito existente na ontologia. Este grau determina qual a probabilidade do conceito estar presente na mensagem sendo que pode ser determinado

um limiar (um grau de similaridade mínimo) abaixo do qual é improvável determinar que um conceito esteja presente. No sistema este limiar é estabelecido por um especialista antes do início de uma sessão no *chat*.

Este método de *text mining* é utilizado tanto para identificar conceitos nas mensagens do *chat* quanto para indexar os documentos eletrônicos da Biblioteca Digital. No caso dos documentos, poderão ser associados vários conceitos . Já no caso das mensagens, apenas o conceito que tem maior grau é associado à mensagem. Se dois conceitos são identificados com o mesmo grau, e se existe entre eles uma hierarquia (pai e filho), o conceito mais específico é utilizado. Caso o empate exista entre conceitos que estão em um mesmo nível da hierarquia, então o sistema irá optar por um deles.

Por último, cabe ressaltar que termos que não estão presentes na ontologia, que forem identificados nas mensagens são registrados para uma análise posterior.

2.3 A Ontologia

O sistema utiliza uma ontologia de domínio para classificar documentos, para identificar temas nas mensagens e para traçar o perfil dos usuários.

Uma ontologia é uma definição formal e explícita de conceitos (classes ou categorias) e seus atributos e relações [Noy e McGuinnes 2002]. Uma ontologia do domínio – *domain ontology* é uma descrição de "coisas" que existem ou podem existir em um domínio [Sowa 1997].

No sistema proposto, a ontologia é implementada como um conjunto de conceitos em uma estrutura hierárquica (um nó raiz, e nós pais e filhos). Cada conceito tem associado a si uma lista de termos e seus respectivos pesos, que ajudam e identificar o conceito presente nos textos.

Os pesos associados aos termos determinam a importância relativa ou a probabilidade de um determinado termo identificar o conceito em um texto. Cabe ressaltar que a relação entre termos e conceitos é muitos-para-muitos, isto é, um mesmo termo pode aparecer em mais de um conceito com pesos diferentes, e um conceito pode ser identificado por diferentes termos.

A ontologia é usada para identificar assuntos nas mensagens de texto do *chat*, para classificar automaticamente itens na Biblioteca Digital e para relacionar pessoas a determinados conceitos (assuntos/temas) visando identificar quais são as pessoas com determinado grau de conhecimento ou interesse em determinado assunto.

A ontologia é utilizada também para recuperar itens da Biblioteca Digital e para procurar por discussões anteriores relacionadas aos conceitos presentes. Estas pesquisas podem ser feitas fora do *chat*, usando outros módulos do sistema.

Para gerenciar e manter a ontologia foram implementados algumas ferramentas que permitem sua visualização (hierarquia de conceitos, termos associados aos conceitos), a inclusão e a remoção de conceitos e de termos e a modificação dos pesos dos termos associados aos conceitos. A tarefa de manter a ontologia cabe a um grupo de administradores.

Na implementação atual, a ontologia está voltada para a área de Ciência da Computação, sendo os conceitos baseados na classificação da ACM. Nem toda a

classificação foi utilizada, sendo que hoje apenas alguns conceitos vem sendo trabalhados.

Os termos associados a cada conceito e os seus respectivos pesos são determinados mediante um processo que começa pela escolha de textos de treino (aproximadamente 100 em inglês e português) identificados por especialistas como associados a um conceito. A partir destes textos, uma ferramenta de software procura identificar os termos que são mais relevantes para cada conceito, procurando definir um centróide para a classe. Uma revisão dos termos e pesos é feita por especialistas nas áreas relacionadas aos conceitos. Outras ferramentas de software auxiliam nesta revisão, por exemplo, apresentando termos presentes em mais de um conceito e identificando termos comuns a conceitos "pai" e "filho".

Na versão atual, somente uma ontologia para a Computação existe no sistema. Entretanto, outras ontologias podem ser adicionadas. É importante ressaltar que cada Comunidade Virtual deverá ter a sua ontologia cadastrada no sistema.

2.4 A Biblioteca Digital

Pode-se definir uma Biblioteca Digital como uma coleção de recursos digitais, organizados sob uma certa lógica e acessíveis para recuperação sobre uma rede de computadores [Kochtanek; Hein e Kassim 2001].

A maioria das Bibliotecas Digitais atua de forma reativa: os usuários formulam consultas usando os mecanismos de busca e recebem a indicação de conteúdos que potencialmente atendem suas necessidades. Já a Biblioteca Digital proposta para o sistema, ajudará no processo proativo de recomendação. Os itens serão recomendados a partir da análise das mensagens enviadas, sem que seja necessário que o usuário toma alguma atitude.

Todo o conteúdo existente na Biblioteca Digital é previamente indexado. O processo de indexação relaciona cada documento, (artigo ou *site*) a determinados conceitos existentes na ontologia, sendo estabelecido o grau do relacionamento entre cada um dos conceitos e o documento.

O processo de indexação de documentos (artigos e sites) é feito de forma automática, sendo que o método utilizado é o mesmo do módulo de *text mining*. O material pode ser colocado na Biblioteca digital por qualquer membro da comunidade, sendo que esta inclusão pode ser feita fora da sessão do *chat*. Feita a inclusão de um novo conteúdo, este 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 *chat*.

2.5 A Base de Perfis

A base de perfis mantém informações sobre os membros da comunidade. Estas informações correspondem a dados cadastrais (nome, e-mail, empresa), e também informações sobre o seu grau de interesse e/ou conhecimento em determinados assuntos (representados por conceitos presentes na ontologia).

A base de perfis do sistema é construída visando aprimorar o processo de recomendação, apresentando para o usuário somente itens que forem do seu interesse e do seu nível de conhecimento.

A base de perfis pode ser usada como Mapa de Conhecimento ou *Yellow Pages*, que são ferramentas que procuram indicar quem dentro de uma comunidade possui determinados conhecimentos ou habilidades. Estes mapas não armazenam a solução de problemas, mas indicam que certas pessoas possuem determinado tipo de conhecimento, estando potencialmente aptas a auxiliar na solução de algum problema [Davenport e Pruzac 1997]. A base também permite identificar a *expertise* de cada membro (o que cada membro mais conhece) e os especialistas em cada área (quem tem mais conhecimento).

No sistema, o perfil é estabelecido inicialmente no momento do cadastramento do membro da comunidade, sendo associados manualmente alguns assuntos que são do interesse e/ou conhecimento do membro. Cabe ressaltar que o sistema irá recomendar conteúdos mesmo a usuários com perfil inicialmente indefinido.

Como o perfil é dinâmico, o próprio sistema irá acompanhar a sua evolução . Esta evolução é constatada a partir das contribuições de cada membro no sentido de incrementar o conteúdo da Biblioteca Digital (inclusões de novos itens), das recomendações utilizadas (realização de *downloads* de artigos recomendados, por exemplo), e participação em discussões (participação passiva ou ativa, através do envio de mensagens dentro do assunto). Cada tipo de participação gera pontos (tipo recompensa) para o membro da comunidade.

A estratégia de pontuação ainda está sendo avaliada, especialmente no que se refere a como mensurar o aumento de interesse ou de conhecimento do usuário no momento em que este realiza algum procedimento.

2.6 A Base das Discussões Anteriores

Sobre cada sessão de *chat* serão armazenados: os participantes da sessão, as mensagens trocadas, o(s) conceito(s) relacionado(s) a cada mensagem (quando for possível identificar estes conceitos) salientando o grau de afinidade com cada mensagem, as recomendações feitas a cada um dos participantes e o aproveitamento das recomendações (quais usuários abriram determinados documentos durante a sessão).

A finalidade de armazenar estas discussões anteriores está em tentar codificar parte do conhecimento tácito existente [Davenport e Pruzac 1997], acompanhar a participação de cada membro do grupo e além disto avaliar os mecanismos do sistema.

No que se refere à tentativa de codificar o conhecimento tácito, sabe-se que grande parte do conhecimento de uma organização não está codificado ou então não está adequadamente estruturado, não sendo então facilmente recuperado. Dentro de uma Comunidade Virtual muito conhecimento pode ser compartilhado indiretamente entre seus membros através da memória do grupo (registro de ações ou conhecimentos). Outros membros poderão vir a ter as mesmas necessidades ou dúvidas e poderão encontrar soluções armazenadas por outros membros em momentos passados.

Futuramente, os usuários do sistema também receberão recomendações de sessões anteriores *chat* que trataram acerca do mesmo assunto corrente, sendo possível a estes usuários recuperar o conteúdo destas discussões. A existência deste registro

de discussões passadas permitirá ainda que seja avaliado qual o assunto principal de uma sessão no *chat*, quais assuntos são recorrentes, com que freqüência os usuários desviaram-se do assunto, se o assunto tratou de forma genérica uma determinada área ou então se houve algum aprofundamento (o assunto passou de um conceito "pai" para um conceito "filho"). Este tipo de análise será melhor comentada na seção sobre os experimentos realizados com o sistema.

2.7 O Módulo de Recomendação

Na maioria das vezes, os membros de uma Comunidade Virtual têm de utilizar ferramentas de busca para recuperar os conteúdos que estão armazenados em algum repositório comum ao grupo. Assim, a recuperação do conhecimento dá-se sempre de forma reativa: a partir de um consulta formulada pelo usuário, uma ferramenta de consulta procura retornar aquele conteúdo que possa ser útil ao usuário. Em algumas situações é interessante oferecer ao usuário o conteúdo que é potencialmente útil, sem que seja necessário que este usuário tenha de informar explicitamente as suas necessidades.

Com a adoção destes mecanismos de recomendação, cria-se a possibilidade do membro de uma comunidade receber proativamente uma série de sugestões de conteúdo.

O objetivo do módulo de recomendação é oferecer o conhecimento e a informação armazenados nas diversas bases (Biblioteca Digital, Base de Perfis, Base de Discussões Anteriores) durante uma sessão de *chat*.

O módulo entra em ação quando recebe um conceito identificado no módulo de *text mining*. A partir disto, o módulo pesquisa as diferentes bases à procura de itens associados ao assunto (conceito) identificado.

As recomendações feitas não devem interromper a discussão. Em função disto as recomendações aparecem em um quadro separado sendo individualizadas, isto é, cada usuário recebe uma lista diferente de recomendações.

A versão atual do sistema não recomenda:

a) o mesmo item mais de uma vez na mesma sessão;

b) itens da Biblioteca Digital classificados como básicos para membros especialistas no mesmo conceito dos itens (um limiar é utilizado para determinar os especialistas);

c) itens já recomendados. Este critério ainda está sendo avaliado.

A Figura 1 mostra a interface do sistema em uso. O primeiro quadro em cima à esquerda mostra os membros ativos no *chat*. O quadro à direita deste mostra as mensagens enviadas e por quem. As recomendações são feitas no quadro abaixo. À esquerda são mostrados os conceitos da ontologia que foram identificados nas mensagens.

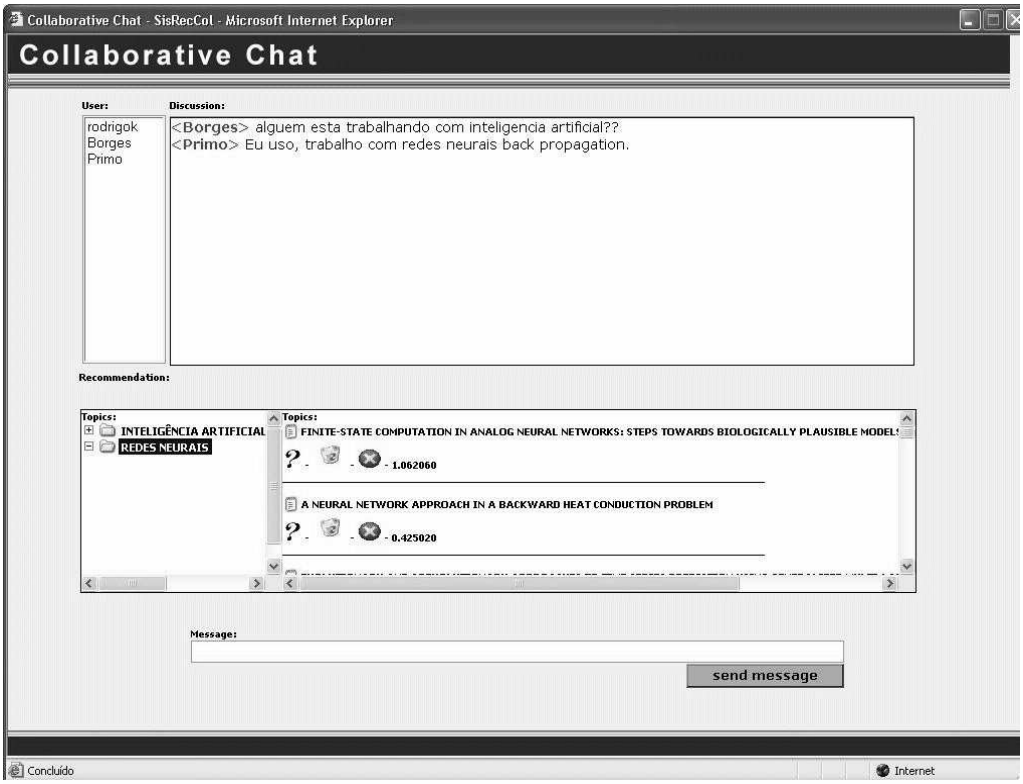

Figura 1. Interface do Sistema

3. Trabalhos Correlatos

Esta seção procura apresentar alguns trabalhos que estão relacionados ao sistema aqui descrito, estabelecendo algumas comparações.

Em primeiro lugar o sistema aqui proposto procura auxiliar no processo de recuperação colaborativa descrito por [Poltrock et al. 2003]. Esta difere da recuperação individual porque envolve comunicação entre os membros de uma comunidade, o compartilhamento de informações recuperadas e a coordenação de atividades entre os participantes. O sistema proposto libera os membros do grupo de terem que procurar informações (documentos). Pessoas com diferentes níveis de conhecimento podem-se ajudar. Quando um especialista fala de um tema novo para um novato, o sistema proposto procura itens na Biblioteca Digital, liberando o especialista da tarefa de relembrar conceitos ou fontes importantes que podem ser consultadas.

O sistema proposto irá, em sua versão final, recomendar especialistas reduzindo o tempo gasto na tarefa de identificar pessoas que possuem o conhecimento necessário. Este fato é destacado em [McDonald e Ackerman 2000] que descreve vários sistemas que procuram auxiliar nesta tarefa, citando inclusive sistemas que partem do princípio que trocas narrativas demonstram *expertise* e competência, o que o sistema proposto procura fazer.

A construção do perfil do membro da comunidade está relacionada a Modelagem do Usuário - *User Modelling*. [Middleton; de Roure e Shadbolt 2001] afirma que a modelagem do usuário é baseada no comportamento - *behaivor-based* ou é baseada no conhecimento – *knowledge based*.

No caso da modelagem baseada no conhecimento, os usuários são associados dinamicamente aos chamados modelos estáticos de usuários – *static models of users*. Já a abordagem baseada no comportamento, parte do princípio que o próprio comportamento forma o modelo sendo então aplicadas técnicas de machine-learning para descobrir padrões úteis de comportamento [Middleton; de Roure e Shadbolt 2001].

Já [Agostini et al. 2003] apresenta o sistema MILK, que usa ontologias para associar itens (documentos, projetos, pessoas) e encontrar projetos similares, pessoas interessadas nestes mesmos temas ou documentos similares. O sistema proposto também utiliza uma ontologia para recomendar documentos e pessoas. As pessoas são associadas a conceitos da ontologia de acordo com suas atividades no sistema (que representa a Comunidade Virtual) e esta associação recebe um grau, indicando a força da relação, ou seja, o grau de expertise da pessoa na área referente ao conceito.

[Terveen e Hill 2001] discutem vários sistemas de recomendação. Por exemplo, o sistema PHOAKS extrai páginas Web que foram mencionadas em mensagens do *newsgroup* Usenet para futuramente permitir a recuperação destas indicações. Já o sistema proposto por Donath et al. (*apud* [Terveen e Hill 2001]) analisa discussões do Usenet e de *chats*, atribuindo-lhes propriedades (por exemplo, número de participações). Isto permite recuperar discussões com certas propriedades. Estas ferramentas são úteis para analisar as mensagens mas não geram recomendações durante a discussão. Neste caso, a função é apenas a de registrar as mensagens para permitir a posterior recuperação de informações.

Já o sistema Tapestry permite que as pessoas registrem suas avaliações sobre certas mensagens postadas em uma discussão. Entretanto, neste caso, os processos de armazenar e recuperar conhecimento acontecem de forma explícita, ou seja, as pessoas decidem realizar o processo e tomam a iniciativa. Neste caso, o sistema automatizado funciona de forma **reativa**, esperando um estímulo da pessoa para realizar as suas ações. O sistema aqui proposta realiza a recomendação de forma de forma proativa.

4. Avaliação do Sistema

O sistema foi avaliado a partir dos resultados de alguns experimentos iniciais. A primeira constatação foi que erros ortográficos, gírias, expressões do domínio e abreviaturas tendem a confundir o método de identificação dos conceitos nas mensagens (*Text Mining*). Um corretor ortográfico e a inclusão na ontologia dos demais termos que puderem ser previstos (os mais comuns) devem minimizar tais problemas.

As mensagens para as quais o sistema identificou conceitos errados foram analisadas (e principalmente os pesos associados a elas). Foi possível notar que o limiar utilizado para corte estava muito baixo. Uma possibilidade é deixar que o grupo de participantes defina este limiar em tempo real (para aumentar precisão ou abrangência). Entretanto, pelos resultados dos experimentos, sugere-se utilizar o limiar 0,01 (lembrando que o grau máximo entre uma mensagem e um conceito é 1). Notou-se que pesos acima de 0,0.. estavam corretos e pesos abaixo de 0,0001 indicavam conceitos erroneamente associados. Pesos com valores na faixa 0,00.. às vezes indicavam conceitos corretos e em outras conceitos errados.

Entretanto, mensagens com peso baixo podem indicar a falta de um conceito específico na ontologia, como aconteceu em alguns casos. Uma idéia seria encontrar as

palavras que aparecem nestas mensagens e passar para um especialista decidir se deve ser criado um novo conceito ou não.

A partir da análise dos conceitos erroneamente identificados, também foi possível notar que havia enganos nos pesos dos termos associados a alguns conceitos na ontologia. Termos de significado muito genérico tendem a induzir a erros se tiverem pesos muito altos (por exemplo: projeto, sistema, técnicas). Os pesos destes termos foram diminuídos manualmente na ontologia. Outra maneira de evitar tal problema é cuidando para que os termos de maior peso em cada conceito estejam na mesma faixa de valor (normalizados).

Outro erro comum foi identificar um conceito "filho" quando deveria ser identificado o conceito "pai" (exemplo: o termo "tabelas" levou ao conceito "projeto de banco de dados" ao invés de "banco de dados"). A solução encontrada foi diminuir nos conceitos "filhos" o peso de termos que são mais genéricos, ou seja, identificam melhor o conceito "pai". Está sendo planejada uma ferramenta que encontra palavras que aparecem em conceitos "pai" e "filho", para serem apresentadas a especialista que, manualmente, poderão modificar os seus pesos.

Um caso especial encontrado a partir dos conceitos errados é o de mensagens com vários assuntos (ex: uma mensagem citava "inteligência artificial", "arquitetura de computadores" e outras sub-áreas). Um termo com peso alto num destes conceitos influencia o conceito final identificado. Uma solução pode ser equiparar os pesos nos conceitos (normalização). Mas a questão principal é que somente um conceito vai ser identificado (resposta final do módulo de *Text Mining*), quando na verdade o correto seria indicar os vários conceitos presentes (lembrando que o módulo de *Text Mining* detecta vários conceitos mas somente identifica como conceito correto o de maior peso). Analisando os pesos de todos os conceitos detectados, pode-se notar semelhança de valores. Isto poderia ser utilizado pelo módulo de *Text Mining* para identificar vários conceitos numa mesma mensagem (quando diversos conceitos forem detectados com peso acima do limiar então é porque realmente existem vários conceitos sendo discutidos na mensagem).

Um problema sério do sistema é que ele não trata o contexto da mensagem, ou seja, uma mensagem é independente de outra, fato que não corresponde à maneira real de as pessoas se comunicarem. Assim, por exemplo, se o grupo estiver falando de métodos de "*machine learning*" e alguém enviar uma mensagem falando de "aprendizagem" (e somente isto), o sistema pode erroneamente identificar nesta última o conceito "Informática na Educação".

Uma solução possível seria utilizar modernas técnicas de Processamento de Linguagem Natural para resolver anáforas, elipses e etc. Entretanto, tal solução tem um custo alto de desenvolvimento. Pensou-se então em outras soluções menos complexas. Uma alternativa seria não aceitar variações bruscas de temas, analisando na hierarquia da ontologia a ordem em que os conceitos vão sendo identificados. Por exemplo, seria permitido mudar para um "primo em segundo grau" mas não ir mais longe que isto. Mudanças bruscas indicariam erro no método de *Text Mining*. Neste caso, o próximo conceito no ranking da mensagem corrente e desde que próximo do conceito identificado na última mensagem seria tido como o correto. Isto, é claro, inibiria os desvios de temas que realmente ocorrem numa discussão. Estudos futuros deverão ser feitos para avaliar se esta prática de mudança para temas distantes é comum ou não.

Outra solução viável para o problema da falta de análise de contexto é somente identificar um conceito após várias mensagens terem sido enviadas e os pesos destas somados estarem acima do limiar (uma mensagem única também poderia indicar um conceito). Pelos experimentos, foi possível notar que esta abordagem é a mais correta (e deverá ser adotada na próxima versão do sistema). Por exemplo, notou-se que termos como “indexação”, “tabelas” e “mysql” apareciam em mensagens separadas. Os pesos individuais dos conceitos identificados não passavam o limiar estabelecido. Entretanto, somando-se os pesos de cada conceito detectado (mais de um por mensagem), seria possível indicar corretamente o conceito “banco de dados”, com peso acima do limiar estabelecido.

Outro caso interessante notado nos experimentos é o das mensagens curtas (com 5 palavras ou menos). Algumas permitiram identificar conceitos corretamente, mas a grande maioria ou não levou a nenhum conceito ou gerou a identificação errada. As soluções seriam utilizar o esquema de soma de pesos das mensagens, como descrito anteriormente (para análise do contexto), ou não analisar este tipo de mensagem (definir um número mínimo de palavras para uma mensagem poder ser avaliada).

A análise dos resultados procurou também identificar padrões nas mensagens que não geraram nenhum conceito (onde não foi identificado conceito nenhum, acima do limiar). A primeira constatação é que, na maioria, eram mensagens com poucas palavras (6 ou menos). Entretanto, mesmo algumas mensagens mais longas não permitiram identificar nenhum conceito. Algumas destas chegaram a ter 20 palavras no total, sendo metade destas palavras não significativas (“*stopwords*”). Pôde-se notar também que foram comuns os erros ortográficos e expressões fora de contexto (como “oi”) ou que retrucavam (“que acham ?”, “por quê ?”).

Por fim, o Sistema descrito neste artigo separa numa lista as palavras utilizadas em alguma mensagem que não estavam presentes na ontologia nem eram “*stopwords*”. Analisando-as, notaram-se expressões que realmente não deveriam estar na ontologia (como “xi”, “haha”, nomes próprios, gírias, erros ortográficos e números), mas também puderam ser identificadas novas *stopwords* (verbos genéricos, por exemplo) e termos bastante significativos (exemplo “distribuídos”) que ficaram de fora da ontologia porque a ontologia foi iniciada a partir de ferramentas de aprendizagem automática (“*machine learning*”) aplicadas sobre textos selecionados para treino. Fica claro assim que a ontologia deve ser revisada e uma das formas é coletar os termos usados pela comunidade e que não se encontram na ontologia.

O método de *text mining* foi testado da seguinte forma: resumos, parágrafos e frases de artigos científicos foram fornecidos como entrada para o módulo de *text mining*, simulando as mensagens de um *chat*. Depois, tomou-se o conceito com maior grau associado ao texto de entrada como sendo o tema identificado. O resultado é que resumos e parágrafos (textos maiores) tiveram 100% de acerto, enquanto que as frases isoladas (mesmas frases destas partes) somente chegaram a uma média de 38% de acerto. A conclusão é que frases maiores permitem uma melhor identificação dos temas. Por isto, a nova versão do sistema irá considerar somente um grupo de mensagens para esta identificação e não mais a mensagem isolada.

Por fim, as recomendações foram avaliadas informalmente por pessoas que utilizaram o sistema. A primeira constatação é que os limiares ainda precisam ser melhor estudados, pois o volume de recomendações apresentadas diminuía a precisão

do sistema. A opção de deixar a seleção do limiar para o usuário ajuda mas ainda permite equívocos caso o usuário não tenha muita experiência no sistema (a dúvida é saber qual o limiar mais adequado para não deixar itens importantes de fora nem mostrar muitos itens irrelevantes).

Também notou-se que itens errados (de outros temas) eram recomendados juntamente com itens corretos. Uma explicação está nos erros mencionados acima na identificação de conceitos nas mensagens e nos problemas da ontologia. Ainda há uma pequena sobrecarga de informações devido ao volume grande de itens recomendados. Isto foi minimizado utilizando-se na interface do *chat* uma estrutura gráfica para os conceitos, separando os itens por conceito (conforme pode ser visto na Figura 2). Futuramente, com a implementação do novo método (análise de temas somente num grupo de mensagens enviadas), o número de recomendações deverá diminuir e sua precisão aumentar.

5. Experimentos

Foram realizados experimentos utilizando-se o sistema proposto. Os membros das comunidades virtuais foram convidadas a entrar no *chat* num horário determinado e realizar uma sessão *online* real. Somente grupos da área de Computação puderam participar devido à ontologia ser nesta área. Estão planejados experimentos em aula, onde professor e alunos poderão discutir assuntos de uma disciplina. Acredita-se que tais discussões serão mais restritas a alguns poucos temas da ontologia.

A análise da ordem em que os conceitos vão sendo identificados permite descobrir como a discussão se desenrolou, ou seja, a rota da discussão pelos conceitos. Num dos experimentos, pôde-se notar que ora a discussão "subia" para um conceito "pai" (ex: inteligência artificial) e ora "descia" para um conceito "filho" (sistemas especialistas), indicando algum tipo de especialização na discussão. Em outro experimento, num dado momento da discussão, surgiu um tema bastante diferente (distante na hierarquia) dos temas sendo discutidos. Analisando manualmente esta mudança, verificou-se que realmente procedia e que o membro do grupo que enviou a mensagem descobriu a possibilidade de integrar uma área distante.

Esta análise de rota também permite avaliar os desvios do tema central. Num dos experimentos, notou-se que o tema central aparecia regularmente entre as mensagens, ou seja, pode-se concluir que o grupo discutia assuntos paralelos mas sempre alguém retornava o tema central. Já em outro experimento, o tema central estava dominando o início da discussão mas não mais apareceu no final desta, sendo que as últimas mensagens foram dominadas por um tema "primo" na hierarquia.

A análise estatística dos temas identificados nas mensagens permite descobrir conhecimento interessante em cada comunidade através de suas discussões *online*. Por exemplo, é possível determinar que temas tiveram maior número de mensagens associadas. Isto pode identificar o tema principal da discussão, os temas periféricos e os temas que não tiveram tanta importância. Num dos experimentos, notou-se um número grande de mensagens sobre um tema não muito próximo dos temas mais discutidos. Analisando-se manualmente as mensagens referentes a este tema "estranho", foi possível verificar que a discussão levou a um assunto diferente do objetivo da sessão e que, pelo número de mensagens, tinha uma importância alta, por isto merecendo ser discutido naquele momento.

As estatísticas também permitem saber o grau de atividade ou envolvimento de cada membro nas discussões. Uma lista com os membros e o número de mensagens enviadas por eles para cada tema é um dos resultados estatísticos que o sistema gera. Isto permite também descobrir o tipo de interesse de cada membro. Em outro experimento, foi possível descobrir que um membro do grupo foi o autor de várias mensagens de um mesmo tema, mas que este tema foi pouco discutido pelo grupo (pelos demais membros). A explicação é que o grupo não seguiu as sugestões daquele membro ou não deu muita importância a suas colocações.

6. Conclusão

Este trabalhou apresentou um sistema que realiza recomendações a partir dos assuntos identificados nas mensagens de um *chat*. O sistema procura aprimorar o processo relacionado à aquisição do conhecimento em Comunidades Virtuais, pois permite que sejam conhecidas mais rapidamente as necessidades dos usuários. Outro ponto a destacar é o relacionado ao maior conhecimento que pode se ter do perfil e dos interesses dos membros de uma comunidade virtual mediante a análise das mensagens trocadas por eles. É possível também aprimorar o processo de determinação de autoridades em determinados assuntos (membros da comunidade com determinado nível de conhecimento em determinados assuntos).

Conforme descrito anteriormente, o método usado para identificação dos assuntos contidos nas mensagens vem sendo testado e aprimorado de forma a aumentar a precisão na identificação dos assuntos tratados. Dentre estes aprimoramentos está a tentativa de não apenas analisar cada mensagem individualmente mas também realizar a análise de um grupo de mensagens. Outro ponto que vem sendo discutido está relacionado a criação da ontologia. O processo utilizado ainda requer um grande esforço especialmente por parte de especialistas na área no sentido revisar os termos relacionados a cada um dos conceitos.

Finalmente, a ferramenta apresentada procura contribuir com o processo de integração regional facilitando o acesso a conteúdo digital relevante, bem como o mapeamento de profissionais com determinados interesses. A ferramenta poderá ser utilizada em cursos à distância ou para reuniões virtuais entre profissionais com interesses em comum.

Agradecimentos

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

Referências

Agostini, A. et al. (2003) Stimulating knowledge Discovery and sharing. Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work. Sanibel Island, USA, p.248-257.

Davenport, T. H. e Pruzac, L. (1997) “Working knowledge – How organizations managewhat they know”, Harvard Business School Press, Harvard.

Kochtanek T.R.; Hein K.K. , Kassim A. R. C. (2001) “A digital library resource Web site: Project DL”, Online Information Review, 25(1).

Lewis, D. D. (1998) “Naive (bayes) at forty: the independence assumption in information Retrieval”, in: Proc. European Conference on Machine Learning, Lecture Notes in Computer Science, v.1398, Springer, Berlin, p. 4-15.

Loh, S. ; Wives, L. K.; Oliveira, J. P. M. (2000) “Concept-based knowledge discovery in texts extracted from the Web”, ACM SIGKDD Explorations 2 (1), p. 29-39.

Maes, P. (1994) Agents that reduce work and information overload. Communications of the ACM, v.37 p.31-40.

McDonald, D.W. e Ackerman, M.S. (2000) “Expertise recommender: a flexible recommendation system and architecture” in Proc. ACM Conf. on Computer Supported Cooperative Work, Philadelphia, p.231-240.

Middleton, S. E; de Roure, D. C.; Shadbolt, N. R. (2001) “Capturing knowledge of user preferences: ontologies in recommender systems”. In Proceedings First International Conference on Knowledge Capture, p. 100-107, Victoria, British Columbia, Canada.

Nonaka, I. e Takeuchi T. (1995) “The knowledge-creating company: how japanese companies create the dynamics of innovation”, Oxford University Press, Cambridge.

Noy N. F. e McGuiness, D. L. (2002) “Ontology Development 101: a guide to creating your first ontology”. Disponível em http://protege.stanford.edu/publications/.

Palazzo, L. A. M. e Costa, A. C. R. (2000) ‘Modelos Proativos na Educação Online”. I Simpósio Catarinense de Educação. Itajaí.

Poltrock, S. et al. (2003) “Information seeking and sharing in design teams” IN: Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work. Sanibel Island, USA, p.239-247.

Ragas H. e Koster, C. H. A. (1998) “Four text classification algorithms compared on a Dutch corpus”, in: Proc. SIGIR’98 International ACM-SIGIR Conference on Research and Development in Information Retrieval, ACM Press, Washington, p. 369-370.

Resnick, P. e Varian, H. (1997) “Recommender Systems” Communications of the ACM v.40 p.56-58.

Rocchio, J. J. (1996) “Document retrieval systems - optimization and evaluation”, Ph.D. Thesis, Harvard Computation Laboratory, Harvard University, Report ISR-10 to National Science Foundation.

Sowa, J. F. (1997) “Building, sharing, and merging ontologies”, AAAI Press / MIT press, pages 3-41.

Terveen, L. e Hill, W. (2001) “Beyond recommender systems: helping people help each other”, in: J. CARROLL, ed., Human computer interaction in the new millennium, Addison-Wesley.