

Sistema de Recomendações em Discussões Eletrônicas – Apoio à Construção do Conhecimento em Comunidades Virtuais de Aprendizagem

Adriana Justin Cerveira Kampff^{1,2}, Stanley Loh^{1,3}, Daniel Lichtnow³, José Palazzo Moreira de Oliveira⁴

¹Universidade Luterana do Brasil (ULBRA) – Faculdade de Informática - R. Miguel Tostes, 101 – Canoas/RS – CEP 92420-280

²Universidade Federal do Rio Grande do Sul (UFRGS) - Programa de Pós-Graduação em Informática na Educação

³Universidade Católica de Pelotas (UCPEL) – Escola de Informática - Rua Félix da Cunha, 412 – Pelotas/RS – Brasil – CEP 96010-000

⁴Universidade Federal do Rio Grande do Sul (UFRGS) – Instituto de Informática - Av. Bento Gonçalves, 6500 – Porto Alegre/RS – Brasil

adriana@maristas.org.br, sloh@terra.com.br, lichtnow@ucpel.tche.br,
palazzo@inf.ufrgs.br

Resumo. A Internet possibilita a criação de Comunidades Virtuais de Aprendizagem, propiciando compartilhamento e construção de conhecimento entre pessoas distantes fisicamente. Embora existam várias tecnologias para apoiar essas comunidades, o conhecimento compartilhado e gerado por meio delas poderia ser melhor aproveitado. Este artigo descreve o sistema SisRecCol, que, seguindo uma perspectiva sócio-interacionista, identifica os assuntos abordados em uma sala de discussão na Web (chat room) e, a partir disso e dos perfis dos interagentes, realiza recomendações individualizadas de informações (textos ou discussões anteriores) ou de usuários mais experientes (especialistas), incidindo na zona de desenvolvimento proximal de cada sujeito (Vygotsky, 1984). Posteriormente, a análise estatística das discussões permite avaliar o conhecimento da comunidade como um todo e de seus membros individualmente.

Abstract. Internet enables the construction and sharing of knowledge without physical frontiers. Despite all the technologies available for online discussions, the knowledge shared through electronic media could be better utilized. Following the socio-interactionist theory of Vygotsky (1984), this paper presents the SisRecCol system, that analyzes the themes discussed in a chat room and then recommends information sources in a personalized way, according to the profile of each participant. Through the recommendations, the system intends to complement the knowledge of each individual, reaching what Vygotsky called zone of proximal development. After the discussion session, statistical analyses allow evaluating the discussion and thus the knowledge of the whole community and of each member.

1 Introdução

É crescente o número de ambientes informatizados que procuram permitir o compartilhamento e a construção de conhecimento de forma coletiva. A difusão mundial da Internet e dos ambientes *Web* tem estimulado a criação de um grande número de Comunidades Virtuais de Aprendizagem, buscando reunir pessoas que, embora geograficamente distantes, tenham interesses comuns. As Comunidades Virtuais trocam informações e aprofundam discussões, através da inclusão de referências sobre os temas enfocados, opiniões complementares aos assuntos abordados e do suporte a uma significativa diversidade de experiências relacionadas. As discussões via fórum e salas eletrônicas proporcionam trocas contextualizadas entre os seus membros e são dois dos elementos mais utilizados para apoiar as atividades de aprendizagem.

Este artigo apresenta o sistema SisRecCol, um sistema para recomendações em tempo real em discussões eletrônicas. O sistema consiste de uma *sala eletrônica de discussão (chat room)* que, além de permitir a troca de mensagens entre seus usuários, oferece dinamicamente recomendações que estimulam o estudo e a ampliação dos conhecimentos dos participantes. Para tanto, o sistema identifica os assuntos das mensagens trocadas e faz recomendações personalizadas para cada membro da discussão, apontando conteúdos, artigos, *sites* Web, mensagens de discussões anteriores ou usuários mais experientes no tema em discussão. O objetivo do sistema é favorecer a aprendizagem coletiva (nível social ou interpessoal) e, ao mesmo tempo, as aprendizagens personalizadas (nível individual ou intrapessoal).

O sistema baseia-se em uma abordagem sócio-interacionista. A teoria sócio-histórica de Vygotsky (1984) caracteriza a elaboração do conhecimento como construção coletiva, ou seja, é na interação das pessoas, em um dado contexto social e cultural, que os membros de um grupo se apropriam, internalizam e geram novos conhecimentos. Segundo Vygotsky, a interação social é a base da aprendizagem, sendo que o desenvolvimento intelectual aparece primeiro no nível social (interpessoal) e depois no nível individual (intrapessoal). Sob esta ótica, o processo de comunicação é fundamental para que a aprendizagem se efetive, sendo a linguagem um signo mediador que permite analisar, abstrair e generalizar, e portanto estabelecer conceitos e categorizá-los. Através da linguagem ocorrem as trocas entre os membros de um grupo e desses com indivíduos mais experientes, que indicam possibilidades para a resolução das situações que se apresentam, atuando como mediadores no processo de apropriação de novos conhecimentos.

Nesse sentido, Vygotsky (1984) coloca que cada sujeito apresenta um nível de conhecimento real (aquilo que é capaz de fazer sozinho, que já domina) e um nível de conhecimento potencial (aquilo que é capaz de fazer com a ajuda de outros elementos do grupo). A diferença entre esses níveis é chamada de zona de desenvolvimento proximal. O sistema SisRecCol busca exercer o papel de mediador entre os sujeitos em discussão, considerando o contexto da discussão e as características individuais, para sugerir materiais complementares, de esclarecimento ou para o aprofundamento dos assuntos em questão, buscando ampliar as zonas de desenvolvimento dos aprendizes e, portanto, permitir que novas aprendizagens aconteçam.

Além disto, o SisRecCol possibilita um tratamento estatístico após a realização das discussões, permitindo analisar as participações e os temas discutidos. Desta forma,

é possível descobrir o tópico central da discussão estabelecida, os temas periféricos, a ordem em que os temas foram discutidos, a abrangência (diversidade de temas) e a profundidade (nível de detalhe, temas gerais em relação a temas específicos). Esta análise posterior também permite avaliar a *expertise* do grupo e de seus membros, como por exemplo, quem mais participou (através do número de mensagens enviadas por cada participante), quais as áreas de interesse de cada um (identificando o assunto de cada mensagem) e que áreas mereceram maior ênfase do grupo em cada discussão.

2 Descrição do Sistema SisRecCol

O objetivo principal do sistema SisRecCol é apoiar o processo de construção do conhecimento dos usuários que o utilizam como ambiente de discussões eletrônicas em grupo. A Figura 1 apresenta a arquitetura do sistema. Há um módulo de *chat*, onde a discussão acontece, um módulo de *text mining* para analisar as mensagens enviadas e a discussão toda, um módulo de recomendação que determina que tipo de conteúdo oferecer a cada participante individualmente, e módulos de suporte, como a Biblioteca Digital (com itens a serem recomendados), um histórico de discussões passadas, uma base de perfis dos usuários e uma ontologia de domínio, que é utilizada para classificar itens da biblioteca, mensagens, discussões e áreas de interesses dos usuários. Os módulos serão descritos em detalhe nas subseções seguintes.

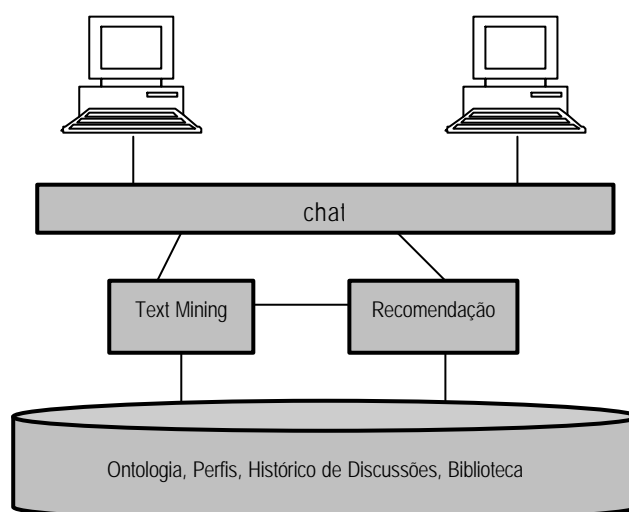


Figura 1: Arquitetura do Sistema

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 cadastrar-se para utilizar o sistema. Isto é necessário para que seja possível criar e manter o perfil do usuário.

2.1 O Módulo de *Text Mining*

Este módulo identifica os assuntos tratados a partir da análise dos termos presentes nas mensagens e da comparação destes com conceitos presente em uma ontologia de domínio. O método utilizado neste sistema foi apresentado em (Loh et al., 2000) e

utiliza técnicas probabilísticas para análise de linguagem natural. Para evitar ambigüidades no momento de identificar o assunto em discussão, um grupo de mensagens é analisado para formar o contexto da discussão. Os termos identificados nas mensagens e que não estejam presentes numa ontologia associada ao domínio da discussão são registrados para uma análise posterior. Um módulo corretor ortográfico é utilizado para corrigir as mensagens antes de sua análise. Feita a identificação do assunto, este é passado para o módulo de recomendação.

2.2 A Ontologia de Domínio

O sistema utiliza uma ontologia de domínio para identificar os temas nas mensagens, classificar os itens da biblioteca digital, classificar as mensagens de discussões anteriores, traçar o perfil dos usuários e relacionar pessoas a determinados conceitos, visando identificar quais são as pessoas com determinado grau de conhecimento em determinado assunto.

Uma ontologia é uma definição formal e explícita de conceitos (classes ou categorias) e seus atributos e relações (Noy & McGuinness 2002). Uma ontologia do domínio (*domain ontology*) é uma descrição de “coisas” que existem ou podem existir em um domínio (Sowa, 2002) e contém o vocabulário relacionado ao domínio em questão (Guarino, 1998).

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 a identificar os conceitos nos textos. Os pesos associados aos termos determinam a importância relativa ou a probabilidade de um determinado termo identificar o conceito em um texto. Os termos associados a cada conceito e os seus respectivos pesos são determinados mediante um processo de aprendizado supervisionado de máquina (*supervised machine learning*), que começa pela escolha por especialistas de textos de treino para cada conceito e termina com um software identificando os termos mais significativos associados ao conceito.

2.3 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 et al., 2001). A Biblioteca Digital implementada no sistema armazena documentos eletrônicos na forma de arquivos eletrônicos (pdf, doc, rtf, txt, html) e apontadores para sites da Web. Este conteúdo é previamente indexado de forma automática pelo mesmo método de mineração de dados textuais utilizado para identificar conceitos nas mensagens. O objetivo da mineração é relacionar cada item aos conceitos existentes na ontologia, sendo estabelecido também o grau de relacionamento entre o item e cada conceito.

Os materiais (documentos ou sites) podem ser cadastrados na Biblioteca digital por qualquer membro da comunidade, sendo que esta inclusão pode ser feita fora da sessão de discussão. Feita a inclusão de um novo conteúdo, o sistema faz a indexação do material e aguarda a aprovação do administrador do sistema para disponibilizá-lo na biblioteca. Uma vez aprovado, o documento passa a estar disponível para utilização, podendo ser recomendado durante as sessões seguintes.

2.4 A Base de Perfis

A base de perfis mantém informações sobre os membros da comunidade que utilizam o sistema. Estas informações correspondem a dados cadastrais (nome, e-mail, etc.) e também a 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 adequados ao seu nível de conhecimento.

A base de perfis também pode ser usada como um Mapa de Conhecimentos 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 & 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), questão esta que será tratada em artigo futuro.

A construção do perfil do usuário está relacionada à modelagem deste (*user modelling*). Middleton et al. (2001) afirmam que a modelagem do usuário é baseada no comportamento deste (*behavior-based*) ou no seu conhecimento (*knowledge-based*). No caso da modelagem baseada no conhecimento, os usuários são associados aos chamados modelos estáticos de usuários. Já a abordagem baseada no comportamento, parte do princípio que o próprio comportamento forma o modelo, sendo então aplicadas técnicas de inteligência artificial (*machine learning*) para descobrir padrões úteis de comportamento.

No SisRecCol, o perfil é estabelecido inicialmente no momento do cadastro do membro da comunidade, através da indicação direta de alguns assuntos que são do interesse e de conhecimento do usuário. Cabe ressaltar que o sistema irá recomendar conteúdos mesmo a usuários com perfis inicialmente indefinidos. Como o perfil é dinâmico, o próprio sistema irá acompanhar sua evolução através das contribuições e participações de cada membro, seguindo a modelagem baseada no comportamento. Cada tipo de participação gera pontos (um tipo de recompensa) para o membro da comunidade e estes pontos são relativos ao assunto em questão. Consideram-se participações válidas: incrementar o conteúdo da Biblioteca Digital (inclusões de novos itens - *upload*), utilizar recomendações (abrir documentos ou sites ou recuperar artigos recomendados) e participar de discussões (participação passiva ou ativa, através do envio de mensagens dentro do assunto). Ainda está em estudo o sistema de pontuação (recompensa).

2.5 A Base de Discussões Anteriores

Cada sessão da sala de discussão é armazenada, ficando associados dados como a data da sessão, os seus participantes, as mensagens enviadas e por quem, o(s) conceito(s) relacionado(s) a cada mensagem (quando for possível identificar estes conceitos), as recomendações feitas a cada um dos participantes e o aproveitamento das recomendações (quais usuários abriram determinados documentos durante a sessão).

Tais registros podem ser utilizados para analisar estatisticamente cada sessão, permitindo descobrir: o assunto principal - aquele que teve maior número de mensagens enviadas; assuntos periféricos - outros assuntos discutidos; o grau de abrangência da discussão - número de conceitos diferentes presentes e suas relações hierárquicas, principalmente conceitos que não têm o mesmo pai; a profundidade da discussão - a que nível na hierarquia de conceitos (descrita na ontologia) chegou a discussão, avaliando os conceitos identificados e suas relações pai-filho; a ordem em que os assuntos foram discutidos - a ordem temporal em que foram identificados, permitindo avaliar se a discussão desviou-se frequentemente ou não, se aprofundou-se ou não, se começou num tópico e terminou noutra ou voltou ao inicial, etc.

2.6 O Módulo de Recomendação

Na maioria das vezes, as pessoas utilizam ferramentas de busca para recuperar conteúdos armazenados em sistemas de computação. Desta forma, a recuperação do conhecimento é feita de forma reativa, pois é necessária uma ação específica do usuário. O interessante é poder oferecer a uma pessoa um conteúdo que lhe é potencialmente útil sem que seja necessário que esta pessoa tenha que informar explicitamente as suas necessidades. Com a adoção de mecanismos de recomendação, cria-se a possibilidade de uma pessoa receber proativamente sugestões de conteúdos ou fontes de informação. O módulo de recomendação é um *software* que recebe um conceito identificado pelo módulo de text mining e oferece conteúdos relacionados ao tema discutido, durante uma sessão na sala de discussões.

As recomendações incluem itens existentes na Biblioteca Digital (documentos eletrônicos e *sites* da *Web*), autoridades no assunto (pessoas com *expertise* acima de um certo limiar no assunto sendo discutido, identificadas na Base de Perfis) e discussões anteriores que versaram sobre o mesmo assunto identificado (armazenadas na Base de Discussões Anteriores). Ao fazer uma recomendação, o sistema irá verificar o perfil de cada usuário participante da sessão, tentando determinar quais os conteúdos relacionados ao assunto que devem ser recomendados a cada um deles.

As recomendações não devem interromper a discussão. Em função disto, as mesmas aparecem em um quadro separado. Ressalta-se que são individualizadas, isto é, cada usuário recebe uma lista diferente de recomendações. O sistema não recomenda o mesmo item mais de uma vez na mesma sessão, nem itens já acessados pelo usuário. Além disto, o usuário pode marcar itens para não serem mais recomendados em sessões futuras. Há diferentes tipos de recomendações, geradas a partir de técnicas diferentes, que não serão detalhadas neste artigo.

A figura 2 apresenta a tela principal do sistema, onde aparecem a área de usuários participantes da discussão (em cima à esquerda), a área de mensagens (em cima à direita), a linha para envio de mensagens (logo abaixo da área de mensagens) e a área de recomendações, dividida por temas (topics) e com os documentos sugeridos.

3 Experimento e Avaliação da Proposta

Atualmente o SisRecCol utiliza somente uma ontologia de domínio, voltada para a área de Ciência da Computação, como projeto piloto. A ontologia implementada representa um subconjunto de conceitos baseados na classificação da ACM e abrange um vocabulário de mais de 3 mil palavras (incluindo termos em português e inglês).

O método para identificar conceitos nas mensagens da sala de discussão foi avaliado formalmente, tendo-se medido que a precisão na identificação de conceitos ultrapassou os 80% de acerto.

Para avaliar o uso do sistema para o fim proposto, foram realizados experimentos com estudantes de diferentes disciplinas de cursos de Computação. Os alunos de cada disciplina foram convidados a entrar na sala eletrônica de discussão num horário determinado e realizar uma sessão.

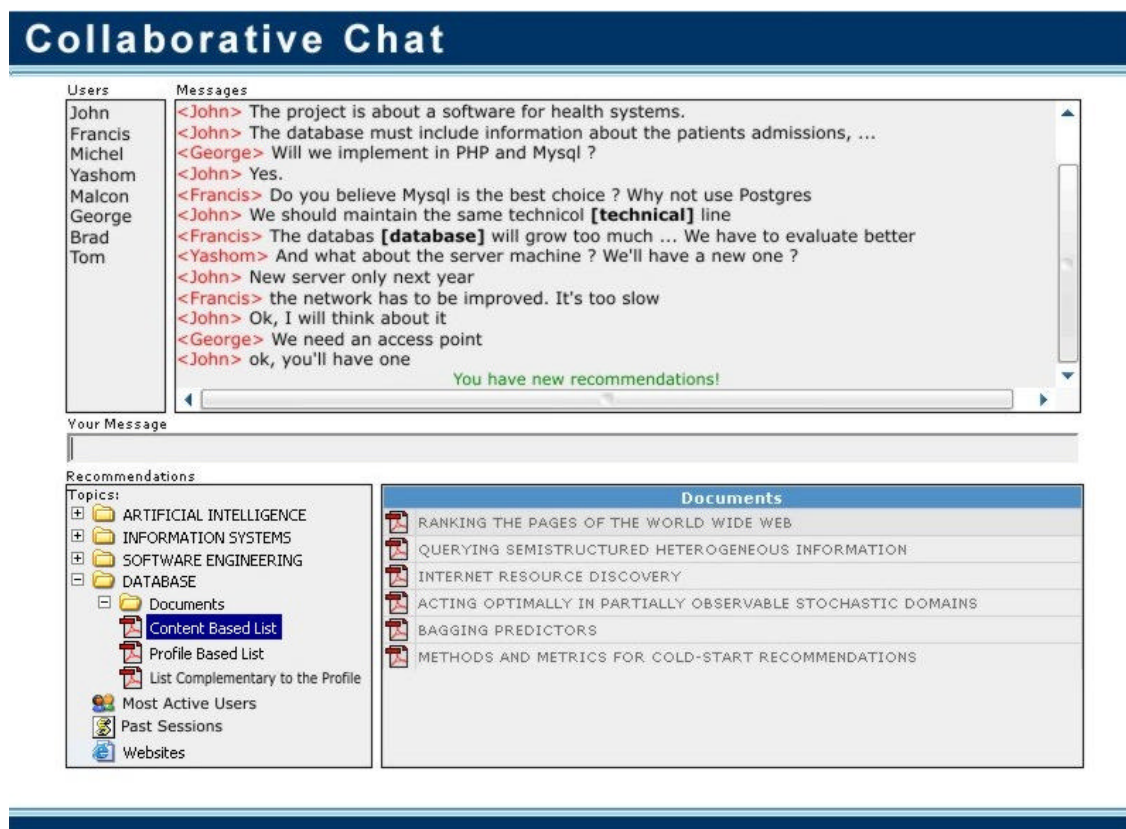


Figura 2: Tela Principal do Sistema

Foi possível notar que o sistema recomenda realmente itens dentro do assunto sendo discutido e isto possibilita ao usuário complementar seu conhecimento, lendo o material recomendado durante a sessão ou posteriormente. Entretanto, a qualidade das recomendações está mais associada à qualidade dos itens presentes na Biblioteca Digital.

A análise da ordem em que os conceitos vão sendo identificados, permitiu 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” (conceito “inteligência artificial”) e ora “descia” para um conceito “filho” (“sistemas especialistas”), indicando algum tipo de especialização na discussão. Esta análise de rota também permitiu 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 ao 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.

Contando-se o número de mensagens associadas a cada assunto, é possível verificar qual foi o assunto mais discutido (aquele com maior número de mensagens) e este então pode ser considerado o tema central. Da mesma forma, pode-se considerar temas periféricos aqueles assuntos que tiveram um número menor de mensagens. O número total de assuntos discutidos pode indicar a abrangência da discussão (diversidade de temas). Já a análise da hierarquia dos temas discutidos pode revelar a profundidade da discussão, ou seja, analisando-se os temas “pais”, “filhos” e “netos”, pode-se saber se a discussão ficou a nível geral ou se chegou a temas mais específicos.

As estatísticas também permitiram 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. Este tipo de análise também permite inferir a *expertise* do grupo como um todo, ou seja, que áreas são de maior conhecimento do grupo e que sub-áreas de um assunto não foram discutidas.

4 Conclusões

Este trabalho apresentou um sistema que realiza recomendações a partir dos assuntos identificados nas mensagens de uma sala eletrônica de discussão (*chat room*). O sistema procura auxiliar a construção coletiva e individual de conhecimento, bem como permitir a análise do conhecimento de um grupo ou de seus indivíduos e do modo como estes compartilham conhecimento.

Não se tem uma medida formal de quanto as recomendações foram úteis. Apenas, sabe-se que alguns documentos foram recuperados durante a discussão, o que demonstra que as recomendações foram de interesse e desta forma podem completar o conhecimento individual. Com um maior número de sessões de um mesmo grupo e um conjunto maior de documentos disponíveis na Biblioteca Digital, espera-se aumentar o fator de utilização das recomendações.

Quanto ao conhecimento coletivo, pôde-se observar que o SisRecCol é uma ferramenta útil para troca de conhecimentos *online*, principalmente para pessoas distantes fisicamente. Além disto, a possibilidade de rever a discussão posteriormente ajuda muito o entendimento de tópicos que não puderam ser pensados durante a discussão, principalmente quando há desvios de assuntos. Da mesma forma, parte do conhecimento tácito fica explicitado no registro das discussões e estas podem ser recuperadas posteriormente por atributos (data, participantes, assuntos discutidos).

Com base nos experimentos realizados, é possível dizer que se pode conhecer melhor o perfil e os interesses de uma comunidade virtual e de seus membros individualmente mediante a análise das mensagens trocadas por eles.

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. Esta estratégia também será usada para recomendar especialistas reduzindo o tempo gasto na tarefa de identificar pessoas que possuem determinado grau de conhecimento numa área.

As limitações do sistema estão relacionadas à construção de novas ontologias e ao volume da Biblioteca Digital, para atender comunidades com interesses diversos. O processo de construção da ontologia é demorado e demanda muito esforço por parte de especialistas. Ainda que ferramentas de software estejam sendo usadas, há o trabalho árduo de selecionar documentos de treino para representar cada conceito. Da mesma forma, identificar que conceitos e quem é “filho” de quem na hierarquia são atividades que exigem conhecimento da área sendo modelada. Estudos sobre a construção automática de ontologias a partir de documentos textuais estão sendo realizados para tentar minimizar tal esforço.

Quanto à Biblioteca Digital, o esforço concentra-se em popular a base com documentos úteis e de qualidade. Estão sendo desenvolvidas ferramentas de software que buscarão automaticamente na *Web*, com base na ontologia armazenada, documentos publicados em congressos e periódicos. Outro software fará a extração dos dados do documento (título, autores, resumo, palavras-chave) para que este possa ser cadastrado na Biblioteca.

5 Agradecimentos

O presente trabalho foi realizado com o apoio do CNPq, uma entidade do Governo Brasileiro voltada ao desenvolvimento científico e tecnológico, e FAPERGS (Fundação de Apoio à Pesquisa do Estado do Rio Grande do Sul).

6 Referências

- Davenport, T. H. & Pruzac, L. (1997) “Working knowledge – How organizations managewhat they know”, Harvard Business School Press, Harvard.
- 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*, 25(1).
- 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.
- 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.
- Noy N. F. & McGuinness, D. L. (2002) “Ontology Development 101: a guide to creating your first ontology”. Disponível em <http://protege.stanford.edu/publications/>.
- Sowa, J. F. (2002) “Building, sharing, and merging ontologies”, AAI Press / MIT press, pages 3-41.
- Vygotsky, L. S. (1984) *A formação social da mente*. São Paulo: Martins Fontes.