Formação de Equipes Profissionais através da Avaliação da Similaridade entre Currículos

Roger Leitzke Granada[1], Paulo Roberto Faulstich Rego[1], Stanley Loh[1,2], Daniel Lichtnow[1], Thyago Borges[1], Gabriel Guimarães[1]

[1]Universidade Católica de Pelotas (UCPEL) – Escola de Informática

[2]Universidade Luterana do Brasil (ULBRA) – Faculdade de Informática

sloh@terra.com.br, lichtnow@ucpel.tche.br,

{ rgranada2004, paulo.faulstich, thyago.borges, asukoto }@gmail.com

Resumo. *Este artigo apresenta um método para identificar a similaridade entre pessoas a partir da análise de seus currículos. O objetivo do trabalho é através de uma função de similaridade encontrar pessoas que possam realizar uma terminada tarefa ou projeto. O trabalho compara 3 formas de análise: a análise dos títulos das publicações, a análise das palavras-chave e a análise do texto das publicações. A aplicação deste método permite dar suporte aos sistemas de recomendação para oferecerem fontes de informações mais relevantes aos usuários.*

1 Introdução

Com a competição cada vez mais acirrada entre as empresas, torna-se importante para as organizações descobrirem expertises similares entre os seus colaboradores a fim de encontrar em um menor tempo possível a "equipe" ideal de profissionais para se executar uma tarefa ou projeto.

A formação de equipes pode ser feita a partir do conhecimento que o gerente de um novo projeto têm dos colaboradores, sendo neste caso bastante informal e dependente do conhecimento do gerente. No caso de uma organização possuir uma base de dados contendo o conhecimento e as áreas de atuação dos seus colaboradores, pode-se utilizar-se delas como Yellow Pages (Mapa do Conhecimento). Yellow Pages são ferramentas que auxiliam a identificação das pessoas e suas habilidades e conhecimentos. As Yellow Pages não armazenam a solução dos problemas, mas com elas é possível encontrar pessoas que possuem a solução (DAVENPORT e PRUZAC, 1997).

Para descobrir as áreas de interesse e do conhecimento dos usuários, podem-se utilizar questionários, entretanto esse método pode não apresentar todas as informações relevantes, uma vez que ao preencher o questionário, há possibilidade de ficarem de fora informações importantes a cerca do seu conhecimento. Também pode ser feita a análise manual dos currículos dos colaboradores, entretanto essa análise pode ser lenta e imprecisa

por necessitar da intervenção humana, especialmente se o número de currículos for muito grande. Nesse ponto a análise automática dos currículos é uma alternativa interessante.

O objetivo desse trabalho é apresentar uma proposta para formar “equipes” de colaboradores com expertises similares a partir da aplicação de técnicas de *Text Mining* sobre textos (currículos) em um pequeno espaço de tempo, e assim selecionar um rol de colaboradores que possam realizar certas tarefas e ou projetos.

Este trabalho está organizado da seguinte forma: na seção 2 são mostrados os trabalhos correlatos; seção 3 é mostrado o método utilizado; seção 4 o experimento realizado e sua avaliação e na seção 5 as conclusões e trabalhos futuros.

2 Trabalhos correlatos

Uma das abordagens para recomendação que mais ganha espaço é a filtragem colaborativa. Ela não considera o conteúdo dos itens, mas sim, a similaridade entre pessoas (TORRES, 2004) cita que o nome Filtragem Colaborativa teve origem no sistema taperstry proposto por (GOLDBERG et al., 1992). A idéia básica é selecionar os itens preferidos pelas pessoas que se assemelham ao usuário alvo, considerando que se usuários concordaram sobre determinados itens no passado irão concordar no futuro. ((RESNICK et al., 1994), (SHARDANAND; MAES, 1995), (HILL et al., 1995)). Os sistemas de Filtragem Colaborativa usam o principio do "Word of Mouth", ou seja, do boca-a-boca. Partindo do pressuposto que quando se procura um livro para ler, um filme para assistir, perguntamos a pessoas que possuem nossos mesmos gostos pela sua opinião.

Grande parte dos algoritmos de Filtragem Colaborativa utiliza o método dos vizinhos mais próximos, onde um número de usuários é recuperado baseando-se na similaridade com o usuário alvo. O artigo de (RESNICK et al., 1994) ou de (HERLOCKER et al., 199) mostram de maneira mais aprofundada a implementação de tal algoritmo. Existem dois tipos de filtragem colaborativa usuário-usuário e item-item.

(ADOMAVICIUS e TUZHILIN, 2006), apresentam aproximações usuário-usuário (user-user) baseada em modelos, o sistema pode gerar um modelo individual para cada usuário das publicações associadas a ele(escrita, lida ou citada) por fim estes modelos são comparados para determinar o grau de semelhança entre os usuários, através de técnicas de clustering e Rede Bayesiana. As aproximações baseadas em modelos geram modelos compactos e sofrem menos com o problema de escassez (sparsity problems) que acontece quando existirem poucos artigos comuns avaliados pelos usuários. Um dos pontos discutidos pelo artigo é que os trabalhos analisados alocam o usuário em uma só classe. Esta é uma limitação importante partindo da idéia de que o usuário pode ter diferentes interesses dependendo do seu contexto (por exemplo, interesses no trabalho ou em casa). Os autores sugerem usar técnicas avançadas de definição de perfil baseadas em Mineração de Dados (Data Mining), regras, seqüências e assinaturas.

Existem alguns trabalhos que tentam resolver os problemas de "usuários novos" (new user) e os chamados problemas de escassez (sparsity problems) através de abordagens baseadas em memória (memory-based).

RASHID ET AL. (2002) e SAMPAIO ET AL.(2006), Por exemplo, utilizaram uma abordagem baseada em memória para encontrar usuários similares e assim tratar problemas como "usuários novos" (new user) e os chamados problemas de escassez (sparsity

problems). Estas abordagens diminuíram a carga sobre os usuários em avaliar com qualidade um grande número de itens.

STOILOVA ET AL. (2005) propõe a avaliação de similaridade entre pessoas em função de seus sites favoritos. Sites favoritos são utilizados como fonte de conhecimento para saber o que é importante para as pessoas. Tanto de forma implícita como explícita, os favoritos possuem conhecimento sobre URL's, títulos, estrutura hierárquica, browser, plataforma e data (hora de inclusão ou de acesso). Embora páginas da web que se encontram nos favoritos possam ser utilizadas para representar o perfil do usuário (com uma abordagem baseada em memória), o trabalho citado usa temas de hierarquia em árvores para determinar a similaridade entre usuários (representado por seus sites favoritos).

Outra maneira de se determinar semelhanças entre usuários é examinar suas relações ou comunicações sociais. SPERTUS ET AL. (2005) compara seis medidas distintas de semelhança de recomendação on-line entre comunidades da rede social Orkut. Essas medidas apresentam a interação entre duas comunidades, como, o número de usuários em comum. Embora isto seja uma avaliação item-item, o trabalho citado sugere uma determinação de semelhança entre usuários (avaliação usuário-usuário) através de membros comuns em uma comunidade ou pela distância entre usuários em um gráfico de amizades.

DUMAIS & NIELSEN (1992) falam do problema sob a perspectiva de uma banca avaliadora: Ele/ela necessita encaminhar artigos para revisores. O problema é remido a um cruzamento de dois tipos de objetos: Artigos e Revisores. Artigos são representados por seu título e resumo (armazenado em vetores de termos). O trabalho comparou as diferentes técnicas para representar a expertise e os interesses do revisor; Os autores tentaram usando sobrenomes, palavras-chaves e resumos extraídos de artigos fornecidos pelos revisores como a melhor representação de seu conhecimento. A expertise e os interesses do revisor são representados por um vetor de termos (termos correspondendo a nomes ou palavras-chave ou ainda por outras palavras apresentadas no artigo). Usando-se a "Latent Semantic Indexing" (Semântica de indexação latente) para cruzar vetores de artigos e de revisores, os resumos mostraram-se a melhor alternativa, com um pequeno ganho (apenas 4% melhor que sobrenomes e palavras-chave). Entretanto as técnicas não são melhores que o desempenho humano.

3 Similaridade entre currículos

O objetivo deste trabalho é encontrar "equipes" de colaboradores através da análise da similaridade entre partes do currículo Lattes (palavras-chave e títulos dos artigos publicados) e entre textos completos das publicações. A descoberta da similaridade entre documentos textuais se dá através da comparação entre eles. O método proposto utiliza as publicações dos usuários para representar seus interesses. O trabalho compara partes das publicações das pessoas que estão presentes nos currículos (títulos, palavras-chave e textos completos) para representar os perfis e encontrar a similaridade entre eles.

Para este trabalho optou-se pela medida denominada de Coeficiente de Jaccard, que foi usada para medir a similaridade entre textos com base na presença dos atributos em pelo menos um dos textos considerados.

O método tem seu início com a transformação dos artigos completos do formato PDF para o formato TXT, para a ferramenta utilizada poder ter acesso aos dados, possibilitando o cálculo da similaridade entre. As palavras-chave e os títulos das publicações utilizadas neste trabalho foram extraídos dos currículos no formato Lattes de professores. O Formato Lattes é desenvolvido pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq 2006) e permite a organização da estrutura dos currículos através de tags XML, facilitando com isso a identificação de seu conteúdo. O próximo passo é a retirada do texto das stopwords, isto é, palavras muito comuns e que não contribuem para identificação do contexto de um determinado documento (VANRIJSBERGEN,1979) (KOWALSKI,1997). Para a remoção das stopwords, foi utilizado um arquivo de texto com termos muito comuns (preposições e artigos) em português e em inglês. Após as palavras resultantes no texto são colocadas em um vetor. A similaridade entre dois vetores é calculada pelo número de atributos comuns(C) dividido pelo número total de atributos sem contar as repetições (número de atributos do primeiro vetor(A) mais o número de atributos no segundo vetor(B) menos o número de atributos comuns aos dois vetores(C)), conforme mostrado na Figura 1.

$$Cj = \frac{C}{A + B - C}$$

Figura 1. Fórmula do coeficiente de Jaccard

Para encontrar a similaridade entre textos nesse trabalho foram feitos experimentos que será mais bem exposto na secção 4.

4 Experimentos e avaliações

Para avaliar o método no contexto da identificação de similaridade entre usuários, foram realizados experimentos com as publicações presentes em currículos no formato Lattes de 10 professores da escola de informática da UCPel. Os currículos no formato Lattes foram obtidos com professores da própria Escola de Informática da Universidade Católica de Pelotas. Esses professores foram selecionados e separados em duplas de acordo com a área de interesse que atuam. Pegando-se assim, dois professores de cada área. Esses currículos foram selecionados de forma que tivessem dois professores relativos a cada área que atuam. Assim, poderiam comparar-se as duplas de similaridade. Foram extraídos títulos e palavras-chave das publicações. Todas as publicações foram consideradas e também foram reunidos os textos completos destas publicações.

Foi escolhido o Currículo Lattes por ser um padrão já estabelecido e por ser uma fonte completa de informação contendo dados sobre formação acadêmica, titulações, artigos escritos e área de trabalho.

Os usuários foram separados previamente em duplas de acordo com a área de interesse determinada previamente por eles. Sendo assim, para a função de similaridade alcançar 100% de acerto, as duplas encontradas deveriam corresponder as duplas previamente estabelecidas.

O cálculo de similaridade foi feito comparando os vetores criados de um documento (vetor de palavras) com todos os outros. A comparação que tivesse maior grau de similaridade com o documento analisado era separada, identificando assim uma dupla de similaridades.

O método utilizado para analisar os títulos e palavras chaves foi o mesmo utilizado para analisar o texto completo das publicações conforme descrito anteriormente. Os resultados (maior grau de similaridade) foram avaliados por cada pessoa, sendo que esta indicava qual outra teria maior similaridade com ela. A tabela 1 mostra os resultados obtidos.

Tabela 1: Resultados obtidos pela função de similaridade

	Palavras-chave	Títulos de artigos	Textos dos artigos
Similaridade	100%	90%	60%

Com podemos ver na tabela 1, o resultados dos experimentos envolvendo a análise de palavras-chaves, teve uma melhor performance do que os experimentos que analisaram os títulos dos artigos e os experimentos que analisaram os textos dos artigos publicados. Uma possível explicação para isso é o fato que as palavras-chaves, são palavras que descrevem a área de atuação e ou interesse, sendo assim essas palavras são usadas por todos aqueles que estudam ou trabalham nessa área. Já os títulos dos artigos variam de acordo com o enfoque do trabalho e o tema do congresso onde foi apresentado. Entretanto esse método teve uma performance de 90% de acerto. Dos experimentos realizados, o que teve pior desempenho foi à análise de todo o texto dos artigos. Isso pode ser explicado com o fato de que dependendo do enfoque do artigo ele pode abordar diversos temas que se misturam com outras áreas. Outro fator que pode ter deixado as palavras-chave e os títulos dos artigos com um valor tão alto, é que como foram currículos de professores só da Escola de informática da Universidade Católica de Pelotas, podem ter trabalhos em conjunto desses professores, logo, o título dos artigos e as palavras-chave seriam os mesmos.

Sendo assim o método se mostrou eficiente para encontrar currículos similares analisando as palavras-chaves dos currículos, entretanto mais experimentos devem ser realizados para se poder confirmar o método. Podemos avaliar que é possível encontrar profissionais com aptidões e conhecimentos similares através da análise dos currículos.

5 Conclusões e trabalhos futuros

Este trabalho apresentou uma proposta para identificar "equipes" de colaboradores comparando os seus currículos através de uma função de similaridade. O trabalho concluiu que se pode utilizar currículos para representar perfis de usuários e demonstrou que a análise dos títulos ou das palavras-chave das publicações permite representar os perfis com maior precisão do que somente analisar o textos completos das publicações.

Esta identificação torna-se importante principalmente em organizações, visto que atualmente muitas empresas têm ciência de que o seu maior bem é o conhecimento dos seus funcionários, existindo, porém certa dificuldade de catalogar este conhecimento e de reunir grupos em um pequeno espaço de tempo.

Um trabalho futuro seria definir técnicas que auxiliem em uma melhor definição do perfil. Um exemplo disso é o aprofundamento da análise temporal realizada. A análise temporal torna-se importante, pois para um determinado usuário ser reconhecido como especialista em determinada área, deve-se levar em consideração o período em que trabalhou nesta área. Evitando assim que uma pessoa que tenha mudado de competência (ou de interesses) com o passar dos anos, seja identificado como especialista da área.

Depois de feitos os experimentos, foram obtidas algumas conclusões:

- O melhor desempenho dos testes foi obtido com as palavras-chave, seguido de títulos de artigos e textos dos artigos.
- Não é necessário utilizar os textos dos artigos para representar os perfis dos usuários, pois eles não obtiveram a melhor performance e tem mais sobrecarga de processamento.

O resultado é preliminar e não surpreende muito, pois no experimento podem existir professores que tem publicações em comum, logo os títulos dos artigos e as palavras-chave seriam as mesmas. Além disso, o resultado foi semelhante ao apresentado por BRUTLAG e MEEK (2000), onde os cabeçalhos de e-mail representam tão bem quanto as mensagens do e-mail para uma classificação da mensagem, com a vantagem adicional de reduzir o número de atributos para serem analisados. Uma possível razão para este achado é o texto completo permitir identificar muitos assuntos enquanto títulos e palavras-chaves se concentraram em menos assuntos e mais específicos. O método para identificar assuntos no texto considera muitas possibilidades e isto pode enganar a avaliação da similaridade, partindo do principio que muitos assuntos não-comuns podem aparecer ao comparar dois autores.

Um trabalho futuro seria fazer essa mesma avaliação só que com professores de outras instituições, para que não se tenha no experimento pessoas que tenham publicado em conjunto.

Outra possível causa do desempenho ruim dos textos dos artigos pode ser a falta de métodos avançados como Stemming. Um estudo futuro poderá avaliar se melhora o desempenho da função de similaridade quando aplicada a técnica de Stemming nos textos.

Agradecimentos

Este trabalho é parcialmente financiado pelo CNPq, uma entidade do Governo Brasileiro para o desenvolvimento científico e tecnológico (Project DIGITEX - Editoração, Indexação e Busca em Bibliotecas, número 550845/2005-4), e FAPERGS, Fundação para Apoio à Pesquisa no Rio Grande do Sul (Projeto Rec-Semântica - Plataforma de Recomendação e Consulta na Web Semântica Evolutiva, número 0408933).

Referências

ADOMAVICIUS, G. and TUZHILIN, A. (2005) "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, v. 17, n. 6, p. 734-749.

BRUTLAG, J.D. and MEEK, C. (2000) "Challenges of the email domain for text classification". In: 7th International Conference on Machine Learning (ICML 2000), Stanford University, USA, p. 103-110

CNPq, Conselho Nacional de Pesquisa e qualidade. Disponível pela URL: http://lattes.cnpq.br/

DAVENPORT, T. H. e PRUZAC, L. (1997) "Working knowledge – How organizations Manage what they know", Harvard Business School Press, Harvard.

DUMAIS, S. T. and NIELSEN, J. "Automating the assignment of submitted manuscripts to reviewers". In 15th International ACM SIGIR Conference on Research and Development in Information Retrieval, Copenhagen, Denmark, p. 233-244.

GOLDBERG, D.; NICHOLS, D.; OKI, B. M. and TERRY, D. "Using collaborative filtering to weave an information taperstry". Commun. ACM, New York, NY, USA, v.35, n.12, p61-70, 1992.

HERLOCKER, J.L.; KONSTAN, J.A.; BORCHERS, A. and RIEDL, J. "An algorithmic framework for performing collaborative filtering". In: SIGIR '99 Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval, 1999, New York, NY, USA. Anais... ACM Press, 1999. p.230-237.

HILL, W.; STEAD, L.; ROSENSTEIN, M. and FURNAS, G. "Recommending and evaluating choices in a virtual community of use". In: CHI'95: Proceedings of the sigchi conference on human factors in computing systems, 1995, New York, NY, USA. Anais... ACM Press/Addison-Wesley Publishing Co.,1995. p.194–201.

KOWALSKI, G. "Information Retrieval Systems: Theory and Implementation". Norwell, MA, USA: Kluwer Academic Publishers, 1997.

RASHID, A. M.; MCNEE, S.M.; ALBERT, I.; COSLEY, D.; GOPALKRISHNAN, P.; LAM, S. K.; KONSTAN, J. A. and RIEDL, J. (2002) "On the recommending of citations for research papers".In Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work, p.116-125.

RESNICK, P.; IACOVOU, N.; SUCHAK, M.; BERGSTORM, P. and RIEDL, J.GroupLens: "An Open Architecture for Collaborative Filtering of Netnews". In: Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, 1994, Chapel Hill, North Carolina. Anais... ACM,1994. p.175–186.

SAMPAIO, I.; RAMALHO, G.; CORRUBLE, V. and PRUDÊNCIO, R. (2006) "Acquiring the preferences of new users in recommender systems: the role of item controversy". In Proceedings of the Workshop on Recommender Systems (in conjunction with the 17th European Conference on Artificial Intelligence - ECAI 2006). Riva del Garda, Italy, August 2006, p.107-110.

SHARDANAND, U. and MAES, P. “Social Information Filtering: Algorithms for Automating WordofMouth”. In: ACM CHI’95 Conference on human factors in computing systems, 1995. Proceedings...[S.l.:s.n.], 1995.v.1,p.210–217.

SPERTUS, E.; SAHAMI, M. and BUYUKKOKTEN, O. “Evaluating similarity measures: a large-scale study in the orkut social network”. Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery and data mining KDD '05, August 2005, p.678-684.

STOILOVA, L.; HOLLOWAY, T.; MARKINES, B.; MAGUITMAN, A. G. and MENCZER, F. (2005) “GiveALink: mining a semantic network of bookmarks for web search and recommendation”. In Proceedings of the 3rd International Workshop on Link discovery LinkKDD, August 2005, p. 66-73.

TORRES, R. “Combining Collaborative and Content-based Filtering to Recommend Research Papers” [S.1.].

VANRIJSBERGEN, C. J. “Information Retrieval, 2nd edition”. [S.l.]: Dept. of Computer Science, University of Glasgow, 1979.