Identificação Automática de Expertise Analisando Currículos no Formato Lattes

Thyago Borges[1], Luiz Carlos Ribeiro Junior[1], Gustavo Piltcher[1] Stanley Loh[1,2], Daniel Lichtnow[1], Rodrigo Branco Kickhöfel[1], Cleber Gouvea[1], Ramiro Saldaña[1]

[1]Escola de Informática – Universidade Católica de Pelotas (UCPel)
Rua Felix da Cunha, 402 – 96.010-000 – Pelotas – RS – Brasil

[2]Informática – Universidade Luterana do Brasil (ULBRA) Rua Miguel Tostes, 101, Canoas, RS, Brasil

{thyago,lcr,lichtnow,rodrigok,cgouvea,rsaldana}@atlas.ucpel.tche.br, sloh@terra.com.br ;gustavopil@gmail.com

Abstract. *This paper presents a software system that automatically identifies expertise in personal curriculums, stored in the Lattes format. The identification is made through the extraction of textual information from XML structures used in the Lattes format. Text mining techniques are used to classify the texts according to themes defined in a domain ontology. This process allows identifying user's expertise, that is, competences and areas of interest. Since curriculums in Lattes format are structured by sections (publications, experience, projects, etc), it is possible to identify different areas in different fields of action.*

Resumo. *Este artigo descreve o desenvolvimento de uma ferramenta para a identificação automática da expertise em currículos (currículo vitae) no formato Lattes. Esta identificação se dá através da extração de informações do arquivo XML gerado pela Plataforma Lattes. Uma vez feita à extração, são utilizadas técnicas de Text Mining de forma a fazer a associação dos dados extraídos a uma ontologia de domínio para a identificação das áreas de atuação e expertise dos usuários. Este trabalho esta sendo desenvolvido para dar auxilio ao Sistema de Recomendação para Apoio à Colaboração (SisRecCol), desenvolvido pelo GPSI (Grupo de Pesquisa em Sistemas de Informação) da UCPel.*

1. Introdução

Para que uma organização possa tornar-se mais competitiva dentro do seu ramo de atividade é necessário que ela conheça o perfil de seus funcionários e também dos candidatos a ingressar nela. Para isso, faz-se imprescindível conhecer as competências destas pessoas para uma melhor avaliação de seus potenciais.

Convém ressaltar que com a velocidade com que surgem novas tecnologias e a constante reformulação de conhecimentos já consolidados, torna-se necessário uma maior agilidade na atualização do conhecimento acerca das competências dos colaboradores de uma organização.

A organização deve manter uma base de dados sobre as pessoas e suas áreas de atuação ou conhecimento, que funcione como um Mapa de Conhecimento *(Yellow Pages)*. Estes mapas não armazenam a solução de problemas, mas indicam quem dentro da comunidade possui determinado tipo de conhecimento, estando virtualmente apto a auxiliar na solução de algum problema (Davenport & Pruzac, 1997).

Neste sentido, definir o perfil de um membro de uma organização, reconhecer suas preferências, suas áreas de interesse é uma tarefa importante. Uma forma de obter estas informações é mediante a aplicação de um questionário a ser preenchido pelo membro da organização.

Ocorre, porém que o preenchimento dos dados solicitados no questionário pode não ser preciso. Além disto, é necessário levar em consideração o fato de que o perfil e os interesses de uma pessoa são dinâmicos. Assim, as informações sobre o usuário podem ficar obsoletas rapidamente, sendo necessário que sejam feitas atualizações destas informações sempre que o membro da organização começar a trabalhar em uma nova área ou quando ele aperfeiçoar seus conhecimentos.

Para minimizar este problema torna-se interessante à identificação e atualização automática do perfil e dos interesses dos usuários. É importante que esta identificação possa ser feita a partir de uma fonte de dados que o usuário costume manter atualizada.

Este artigo apresenta um software que tem por objetivo auxiliar na identificação das *expertises* dos usuários, bem como analisar as competências de um grupo. A ferramenta utiliza o arquivo XML gerado pelo Currículo Lattes desenvolvido pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, 2004).

O presente trabalho está sendo desenvolvido para dar suporte a um Sistema de Recomendação para Apoio à Colaboração (SisRecCol), desenvolvido pelo Grupo de Pesquisa em Sistema de Informações – GPSI (http://gpsi.ucpel.tche.br). Este sistema identifica os assuntos discutidos em seções de um *web chat* privado e então recomenda itens de uma Biblioteca Digital, discussões anteriores e autoridades no assunto (pessoas que possuem conhecimento mais significativo no assunto tratado), podendo assim ajudar as organizações na Gestão de Competências.

A seção 2 apresenta alguns trabalhos correlatos. A seção 3 discute a extração dos dados utilizando arquivos XML dos currículos do formato Lattes. A seção 4 descreve as técnicas de *Text Mining* e a utilização de uma ontologia desenvolvida para identificar a *expertises* dos usuários. Na seção 5 são mostrados experimentos da extração de informações de um Currículo Lattes e testes com artigos científicos para identificar o grau de confiabilidade do módulo de *Text Mining*. O artigo termina com a seção 6 apresentando as conclusões e trabalhos futuros.

2. Trabalhos Correlatos

Yimam-Seid e outros (2003) apresentam trabalhos que criam bases de dados das competências dos usuários. Dentre as ferramentas apresentadas está o *Expert/Expert-Locator (EEL)* que constrói o índice da *expertise* dos grupos baseando-se em documentos produzidos pelos membros destes grupos. Também é apresentado o *ContactFinder* que utiliza agentes inteligentes para monitorar discussões relacionadas a um determinado assunto em *e-mails* identificando os usuários que fazem uso do suporte técnico da organização. É também apresentado o trabalho de Schwartz & Wood que utiliza algoritmos heurísticos para identificar através de mensagens de *e-mail* interesses compartilhados entre grupos de pessoas. O sistema *DEMOIR* permite encontrar peritos em um determinado domínio baseado em uma base de dados chamada *trailbase*, que consiste nas URLs visitadas e nas *keywords* utilizadas para pesquisa destas páginas.

McDonald e Ackerman (2000) descrevem vários sistemas que auxiliam na tarefa de identificar pessoas que possuem determinado conhecimento e propõe a arquitetura de uma ferramenta denominada *Expertise Recommender*. Para realizar a identificação destas pessoas, McDonald e Ackerman (2000) partem do princípio que trocas narrativas demonstram *expertise* e competência. São então definidas regras heurísticas para realizar tais identificações num ambiente de projeto de *software*, tais como: quem alterou por último um determinado módulo e quem resolveu um problema semelhante.

Cabe ainda citar Agostini e outros (2003), que descrevem o sistema *MILK*, que usa ontologias para encontrar projetos similares, pessoas que tem interesses similares ou documentos similares. As pessoas são associadas a conceitos de uma ontologia de acordo com suas atividades no sistema e esta associação recebe um grau, indicando a força da relação, ou seja, o grau de *expertise* da pessoa na área referente ao conceito.

Os sistemas descritos procuram, na maioria dos casos, monitorar as ações dos membros de uma organização, o que sem dúvida nenhuma pode ser útil. Ocorre que em muitas situações este monitoramento não pode ser feito, já que em muitas organizações nem toda atividade realizada está registrada e em alguns casos este monitoramento pode implicar em invadir a privacidade (especialmente no caso de monitoramento de e-mails e da navegação do usuário). Além disso, no momento em que um novo membro ingressa na organização não é possível determinar suas áreas de interesse.

3. Sistema para Identificação de Expertise em Currículos

No presente trabalho, a definição automática do perfil do usuário é feita através da extração de informações em textos, obtidas em um arquivo XML gerado a partir das ferramentas disponíveis no Currículo Lattes, componente que integra a Plataforma Lattes desenvolvida pelo CNPq.

O Currículo Lattes foi escolhido por ser um padrão já estabelecido (especialmente na comunidade acadêmica), por ser uma fonte completa de informação contendo dados sobre formação acadêmica, titulações, artigos escritos e área de trabalho. Além disto, pesou o fato de ser possível gerar um arquivo XML com os dados do currículo.

Na ferramenta, em um primeiro momento é feita à extração automática de todas as informações existentes no arquivo gerado pela plataforma Lattes. Levando-se em consideração o fato de se tratar de uma ferramenta que rodará em ambiente *Web* e que a linguagem utilizada no contexto do projeto SisRecCol (descrito na introdução) é PHP, foi necessário buscar recursos oferecidos por esta linguagem para realizar a extração dos dados.

Foram utilizados os métodos oferecidos pela classe XPath (http://sourceforge.net/projects/phpxpath/), que manipula arquivos XML.

Uma vez extraídos as informações do currículo, estes dados são passados para uma ferramenta que aplica uma técnica de *Text Mining*. A técnica utilizada consiste em comparar as palavras extraídas do currículo Lattes com as palavras presentes em uma ontologia de domínio. Detalhes relacionados ao método são fornecidos na seção 3.1.

Terminado o processo é apresentado o conjunto de conceitos (áreas) aos quais o usuário está relacionado. Estes conceitos estão presentes na ontologia e apontam os assuntos ou temas que são representativos dentro de uma organização.

3.1 A Técnica de Text Mining

Para a identificação das áreas de competência e das *expertises* é utilizada neste trabalho uma técnica de *Text Mining* que trabalha examinando o conteúdo de todo o currículo.

Conforme descrito na seção 3, essa identificação é feita pela comparação de termos que aparecem no currículo, com termos que estão relacionados aos conceitos presentes na ontologia do sistema.

Neste sistema esta sendo utilizado um tipo de classificação baseado em técnicas probabilísticas, que foi apresentado em Loh e outros (2000). Com base nos algoritmos de Rocchio e Bayes (Rocchio, 1966; Ragas & Koster, 1998; Lewis, 1998), é usado um algoritmo que representa textos e conceitos através de vetores. Estes vetores são compostos por uma coleção de termos com um peso associado a cada termo. No caso dos textos que compõe os currículos, o peso de cada termo é dado pelo cálculo da Freqüência Relativa de cada termo no currículo. A Freqüência Relativa é o número de ocorrências da palavra no currículo, dividido pelo número total de palavras presentes no currículo. O peso de um termo representa a probabilidade deste termo indicar um determinado conceito presente na ontologia. Na montagem dos vetores são ignorados os termos que não tem muita relevância na identificação dos conceitos ou que aparecem com muita freqüência, como as preposições, artigos e etc. Estes termos são chamados de *stopwords*.

Utilizando uma função de similaridade que calcula a distância entre dois vetores, o método avalia a similaridade entre um vetor (currículo) e um conceito presente na ontologia. A função de similaridade multiplica os pesos dos termos que estão presentes nos dois vetores (currículo e ontologia), sendo que a soma destes produtos, restringida a 1, é o grau de similaridade existente entre o currículo e o conceito existente na ontologia. Este grau determina qual a probabilidade do conceito estar presente no currículo, ou seja, qual a probabilidade de um membro ter afinidade com determinadas áreas representada pelos conceitos existentes na ontologia.

3.2 A Ontologia

A ferramenta utiliza uma ontologia de domínio para classificar currículos, realizando assim a identificação das *expertises* e traçando 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, 2002).

No sistema proposto, a ontologia é implementada como uma estrutura hierárquica, contendo um conjunto de conceitos. Cada conceito tem associado a si uma lista de termos e seus respectivos pesos, que ajudam e identificar o conceito presente nos currículos.

Os pesos associados aos termos determinam a probabilidade deste termo identificar o conceito em um currículo. Cabe ressaltar que um mesmo termo pode aparecer em mais de um conceito com pesos diferentes, e um conceito pode ser identificado por diferentes termos.

Na versão atual, somente uma ontologia para a Computação, baseada na classificação da *ACM – Association for Computing Machinery* (www.acm.org), existe no sistema. Entretanto, outras ontologias podem ser adicionadas. É importante ressaltar que para cada nova área de aplicação torna-se necessário construir uma ontologia no sistema que se adapte aos interesses de cada organização, possibilitando assim a identificação das competências nas diversas áreas do conhecimento.

5. Experimentos

Antes de avaliar a ferramenta com o uso de currículos, foram feitos testes sobre textos de forma a demonstrar a eficiência do método de *Text Mining* utilizado. Feito isto, foram feitos testes sobre um conjunto de currículos obtidos junto a professores de informática.

Os testes procuraram fazer a identificação de áreas da ontologia relacionadas aos textos e aos currículos.

5.1 Avaliação do método de Text Mining

Foram selecionados entre as áreas da Computação 15 artigos científicos presentes na Biblioteca Digital *ResearchIndex* (www.researchindex.com). Destes artigos foram capturado os resumos (*abstracts*), e então simulado uma entrada no módulo de *Text Mining* com a finalidade de saber qual a porcentagem de acerto do método empregado. Foram avaliados tanto os resumos inteiros de cada artigo, como frase a frase pertencentes aos mesmos resumos. Foram consideradas certas as respostas que correspondiam ao conceito em que este artigo estava indexado na Biblioteca Digital *ResearchIndex*. Por exemplo, se o documento pertencia na base de dados à classe "*Artificial Intelligence*", só seria computado como certo se o módulo de *Text Mining* identificasse o conceito "*Artificial Intelligence*". Foram avaliadas 78 frases obtidas a partir dos resumos. Os resultados são apresentados na tabela 1. Já se esperava em textos maiores, como os resumos inteiros, melhores resultados, pois este método identifica melhor o assunto quando existe um maior número de características (termos) presentes.

Tipo de entrada	% de acertos
Resumos	91,66%
Frases dos resumos	60,97%

Tabela 1: Resultados da Avaliação Offline

5.2 Avaliação da identificação de expertise sobre Currículos

Nos experimentos descritos nesta seção procurou-se demonstrar que além de ser possível reconhecer os interesses atuais de um usuário através da extração das

informações do currículo Lattes, também é possível reconhecer as mudanças de interesse ao longo dos anos.

Nos testes iniciais, foram analisados integralmente os currículos de professores da Escola de Informática da UCPel, sendo considerada correta a identificação se o primeiro conceito retornado do módulo de *Text Mining* estivesse de acordo com a principal área de interesse do usuário. Os experimentos revelaram uma porcentagem em torno de 60% de acertos, isto sem qualquer tipo de filtragem das informações contidas nos currículos, ou seja, sendo analisadas todas as informações mesmo algumas pouco relevantes.

É importante ressaltar que um determinado usuário poderá ter afinidade com diferentes áreas. Neste sentido foi utilizado para os experimentos em currículos um limiar de corte, onde foi estabelecido que conceitos retornados do módulo de *Text Mining* com peso abaixo de 0,001..., seriam descartados por acreditar-se que esses são irrelevantes para análises da *expertise* do usuário. Este limiar pode ser alterado sempre que for necessário.

Foram também feitos testes sobre porções do currículo. Assim, foram escolhidas informações referentes a determinados anos para serem analisados. A saber, foram avaliadas publicações, orientações de graduação, apresentações e etc, nos períodos de 1999 e 2003. A idéia de avaliar períodos de tempo do currículo é tentar verificar possíveis mudanças nas áreas de interesse dos usuários.

Alguns dos resultados que foram obtidos a partir do módulo de *Text Mining* do experimento relativo a um dos currículos avaliados são apresentados nas figuras 1 e 2, como forma de ilustrar os testes. Nestas figuras são mostrados os dois primeiros conceitos da ontologia identificados como sendo áreas de atuação do usuário.

Concept	Weight
DATABASE	0.043614
INFORMATION SYSTEMS	0.042611

Figura 1: Resultados da Avaliação do currículo no Formato Lattes de 1999

Concept	Weight
INFORMATION SYSTEMS	0.015119
NEURAL NETWORKS	0.005865

Figura 2: Resultados da Avaliação do currículo no Formato Lattes de 2003

A partir do teste realizado foi possível identificar as áreas de competência do usuário, assim como as mudanças de suas *expertises* em períodos de tempo. Pôde-se constatar que o usuário conforme a figura 1 teve seu perfil no ano de 1999, relacionado a "*Database*", e conforme a figura 2, no ano de 2003, relacionado ao conceito "*Information Systems*", sendo assim constatada uma leve mudança na área de interesse do usuário no referido período de tempo.

Esta identificação em organizações pode significar um diferencial em relação às outras, uma vez que a partir da definição precisa do perfil de seus funcionários é possível avaliar a real necessidade de treinamentos ou uma possível mudança de domínio da empresa utilizando melhor o conhecimento existente no seu quadro de funcionários ou ainda a necessidade de aquisição de novas competências para atender as suas necessidades.

Por último, um fator importante a ser considerado nos resultados é o fato de que os dados relevantes para a análise correspondem aos termos presentes em algumas partes do currículo, por exemplo, nos títulos dos trabalhos publicados. Em função disto, a possibilidade de acerto é diminuída em comparação a testes feitos com textos maiores como o resumo completo, conforme descrito na seção 5.1.

5.3 Publicações X Orientações

Foram feitos também experimentos analisando as informações relativas a publicações e as orientações de forma separada. Estes testes foram realizados para identificar a existência de uma relação entre as publicações e os trabalhos como orientador.

Os testes foram realizados com a metodologia descrita na seção 5.1. Os resultados indicaram que geralmente a área de publicação é a mesma área que o usuário orienta. Esta relação é quase óbvia, porém, torná-la explícita facilita até mesmo a definição de um orientador por parte de um aluno, indicando exatamente a área que cada professor está se dedicando no momento.

6. Conclusões e Trabalhos futuros

Este trabalho apresentou uma ferramenta que auxilia em processos de Gestão de Competências, permitindo identificar as *expertises* de pessoas através da análise das informações extraídas do arquivo XML gerados através da plataforma Lattes, desenvolvido pelo CNPq.

Esta identificação torna-se importante em organizações, visto que atualmente muitas empresas têm ciência de que o seu maior bem é o conhecimento dos seus funcionários, existindo, porém uma certa dificuldade para catalogar este conhecimento e em determinar quem são especialistas em determinadas áreas.

Isto faz da análise dos currículos dos membros de uma organização uma possível aplicação deste trabalho. Com esta ferramenta, poderá ser possível apontar em um dado domínio as competências presentes aos currículos dos membros de uma organização.

A identificação de perfis através da análise do currículo em XML gerado pela plataforma Lattes, se mostrou bastante eficiente durante os primeiros testes, porém, ainda existem alguns aperfeiçoamentos a serem feitos.

Uma das limitações da aplicação da ferramenta pode estar no uso do currículo no formato Lattes, que vem sendo largamente utilizado no ambiente acadêmico, mas não é adotado com freqüência em outros ambientes. No entanto, os métodos aqui descritos poderão ser utilizados em currículos disponibilizados em outros formatos. Neste caso, a principal limitação irá talvez residir na dificuldade de extrair informações de partes do currículo.

Ficou evidenciado que a ferramenta possui uma forte dependência da qualidade da ontologia e que essa merece atenção especial, sendo necessário pensar em formas de facilitar a sua construção.

Na versão atual, parte do processo de extração de parte dos dados do currículo, como a das informações relativas a determinados períodos de tempo, foi realizada de forma manual, devendo este processo ser aprimorado no futuro.

Versões futuras também levarão em conta o fato de que termos que aparecem em determinadas partes do currículo, deverão ter maior peso na identificação do perfil do usuário. No caso da estrutura do currículo Lattes, termos que aparecem no elemento áreas-de-atuação devem provavelmente ter maior peso do que termos que aparecem relacionados ao elemento participação-em-eventos-congressos.

A análise temporal torna-se importante, pois mesmo que um determinado usuário seja reconhecido como especialista em determinada área, deve-se levar em consideração o período em que trabalhou nesta área. Evitando que uma pessoa que tenha mudado de competência (ou de interesses) com o passar dos anos, seja identificado como especialista.

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.

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 managewhat they know", Harvard Business School Press, Harvard.

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.; OLIVEIR, J. P. M. (2000) "Concept-based knowledge discovery in texts extracted from the Web", ACM SIGKDD Explorations 2 (1), p. 29-39.

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.

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/.

ROCCHIO, J. J. (1966) "Document retrieval systems - optimization and evaluation", Ph.D. Thesis, Harvard Computation Laboratory, Harvard University, Report ISR-10 to National Science Foundation.

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

YIMAM-SEID, Dawit; KOBSA, Alfred. Expert Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach. Journal of Organizational Computing and Electronic Commerce, v. 13, n. 1, 2003, p.1-24.