ISSN 1981-2841

IDENTIFICANDO ESPECIALISTAS A PARTIR DA ANÁLISE DOS ARTIGOS DISPONÍVEIS EM UMA BIBLIOTECA DIGITAL[1]

IDENTIFYING EXPERTS USING INFORMATION AVAILABLE IN A DIGITAL LIBRARY

Daniel Lichtnow[2], Gustavo Piltcher[2], Roger Granada[2], Stanley Loh1[2,3] e Paulo Roberto Faulstich Rego[2]

RESUMO

Descreve-se, neste artigo, o uso de uma técnica de *text mining*, a qual visa a identificar especialistas, usando informações disponíveis em uma biblioteca digital. A técnica utilizada associa autores cadastrados em uma biblioteca digital a conceitos, que apresentam áreas do conhecimento, definidos em uma ontologia do domínio. Os resultados mostram-se em um experimento, no qual se discutem as vantagens que se obtêm com o uso desse tipo de técnica.

Palavras-chave: autor, trabalho científico, ontologia.

ABSTRACT

This work describes a Text Mining technique which aims to identify experts through the analysis of articles available in a digital library. The technique used links authors registered in a digital library to concepts that present knowledge fields. The results are shown in an experiment in which the possible advantages obtained are discussed.

Keywords: *author, scientific work, ontology.*

[1] Trabalho apresentado no V SIRC/ RS.

[2] Acadêmicos do Curso de Sistemas de Informação - Universidade Católica de Pelotas (UCPEL).
[3] Acadêmicos da Faculdade de Informática -Universidade Luterana do Brasil (ULBRA).
{paulo.faulstich,rgranada2004,piltcher}@gmail.com, lichtnow@ucpel.tche.br, sloh@terra.com.br

INTRODUÇÃO

A tarefa de identificação de especialistas vem sendo frequentemente abordada dentro da chamada Gestão do Conhecimento – *Knowledge Management*. Uma das soluções propostas consiste na construção das chamadas páginas amarelas – *yellow pages* (DAVENPORT; PRUZAC, 1997), as quais são repositórios que armazenam informações relacionadas às habilidades e ao conhecimentos de um determinado grupo de pessoas. A construção manual de tais repositórios requererá um investimento considerável de tempo. Assim, buscam-se alternativas que minimizem a execução de tarefas relacionadas à construção das Páginas Amarelas e à localização de especialistas.

Ainda, deve ser considerado que nem sempre as pessoas declaram explicitamente suas especialidades, sendo necessária a busca de evidências que as indiquem. Uma alternativa para fazer a identificação de especialistas consiste em analisar as informações presentes em uma biblioteca digital, de forma que se identificam com que áreas um determinado autor tem maior afinidade.

Neste trabalho apresentam-se alguns experimentos que visam a identificação de especialistas, a partir de informações contidas em uma biblioteca digital – a BDBComp – Biblioteca Digital Brasileira de Computação (BDBComp, 2006). Os procedimentos utilizados consistem na aplicação de um conjunto de técnicas de *Text Mining* que tem por objetivo descobrir o grau de afinidade de um autor presente na BDBComp, em relação a uma determinada área/assunto. O trabalho demonstra que existem situações nas quais o uso de técnicas de *Text Mining* produz benefícios em relação às consultas elaboradas sobre base de dados, usando linguagens como a SQL ou XPath.

Na seção dois do trabalho, desenvolvem-se algumas questões relacionadas à descoberta de especialistas, a partir de informações presentes em repositórios como bibliotecas digitais; na seção três, descreve-se o método de *Text Mining* utilizado; na seção quatro, apresenta-se o processo de extração das informações existentes na BDBComp; na seção cinco, demonstram-se os experimentos realizados e, por fim, na seção seis, apontam-se as considerações finais.

TRABALHOS RELACIONADOS

Diversos sistemas que têm por objetivo auxiliar na descoberta de pessoas com determinado conhecimento são propostos. Esses sistemas, referenciados como *Expert Finders, People-Finder Systems* ou *Expert-Finding Systems* (BECERRA-FERNANDEZ, 2000), possuem a finalidade de descobrir especialistas a

partir da análise de fontes como *home pages* pessoais, conteúdos de *chats, e-mails* e documentos escritos e lidos (D'AMORE, 2005).

O *ContactFinder* (KRULWICH; BURKEY, 1996), por exemplo, apresenta-se como um agente inteligente que identifica especialistas, por meio da análise do conteúdo de um *bulletin board message*. Em Steeter e Lochbaum (1988), apresenta-se o *Expert/Expert-Locator* (EEL) ("*Who Knows"*) que procura identificar a *expertise* de grupos de pesquisa mediante a análise de documentos técnicos produzidos por esse grupo.

O sistema apresentado em Becerra-Fernandez (2000) usa *Term Frequency and Inverse Document Frequency* (TFIDF) sobre os documentos produzidos dentro de um grupo. O sistema assume que um autor é especialista em um determinado assunto, se o documento por ele produzido trata desse assunto. Para descobrir um especialista é fornecida ao sistema uma consulta que se vale de um conjunto de palavras-chave, sendo que o retorno consiste em um conjunto de autores, ordenados segundo o grau de afinidade que têm relação à consulta fornecida como argumento.

Como afirmam Becerra-Fernandez (2000), Steeter e Lochbaum (1988) e D'Amore (2005), as técnicas descritas, neste artigo, também procuram identificar especialistas a partir da análise de documentos produzidos por eles. A diferença reside no fato de que se utiliza uma ontologia de domínio em que se presentificam os assuntos aos quais os autores devem ser associados pelo grau de afinidade. Assim, para descobrir os especialistas, basta indicar um determinado assunto presente na ontologia, não sendo necessário o fornecimento de um conjunto de palavras-chave, processo que é feito em Becerra-Fernandez (2000).

A TÉCNICA DE *TEXT MINING* UTILIZADA

Para descobrir o grau de *expertise* de cada autor presente em uma biblioteca digital utilizou-se uma função de similaridade e uma ontologia de domínio. A mesma técnica foi utilizada em outros trabalhos já publicados, como Lichtnow et al. (2004), Ribeiro Jr. et al. (2005) e Loh et al. (2004), sendo baseada em Loh et al. (2000), Rocchio (1996) Lewis (1998). A técnica consiste em um processo de classificação que considera um determinado conjunto de assuntos/conceitos, previamente definidos em uma ontologia de domínio, faz-se, a partir da classificação, a associação de cada autor aos assuntos. Uma ontologia do domínio – *domain ontology* pode ser definida como uma descrição de "coisas" que existem ou podem existir em um domínio *Sowa* (2002).

No presente trabalho, a ontologia consiste em uma estrutura hierárquica de conceitos/assuntos (um nó raiz, e nós pais e filhos), de modo que cada

conceito tem associado a si uma lista de termos com seus respectivos pesos. Os pesos associados aos termos determinam a probabilidade que um determinado termo tem de identificar o assunto. A ontologia utilizada nos experimentos está limitada à área da Ciência da Computação e os conceitos baseados na classificação da ACM. A Figura 1 ilustra a estrutura da ontologia, porém os pesos mostrados são apenas ilustrativos.

A ontologia foi criada por um processo semiautomático, especialistas desenvolveram a hierarquia de conceitos e separaram documentos associados a cada assunto. Aos documentos (cerca de 100 por assunto), utilizou-se um *software* que gerou o vetor inicial de termos associados a cada conceito. Feito isso, necessitou-se da intervenção humana para corrigir alguns erros e resolver algumas ambiguidades, como reduzir o peso de termos genéricos, incluir a variação morfológica dos termos sinônimos, ajustar alguns pesos e incluir alguns termos.

Na criação da ontologia serviu-se de um *software* que normalizou os vetores gerados, ou seja, o peso de cada termo pertencente aos assuntos da ontologia foi recalculado. O objetivo, nesse procedimento, foi evitar que entre os diferentes conceitos/assuntos da ontologia houvesse uma grande diferença entre os pesos atribuídos aos termos. Para exemplificar, um termo com maior valor numérico associado a um determinado conceito/assunto deveria ter um valor similar a um outro termo que tivesse também o maior valor numérico associado a outro conceito.

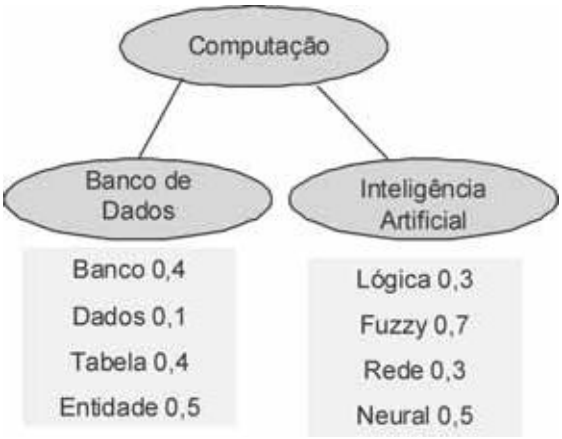

Figura 1. Ontologia.

A similaridade entre um texto (no contexto desse trabalho, o título dos artigos presentes na biblioteca) e um assunto presente na ontologia determina-se pelo

cálculo da distância entre dois vetores, um construído a partir dos termos presentes nos títulos dos artigos de um autor e outro relacionado aos assuntos da ontologia.

Para aplicação da função da similaridade, o primeiro passo consiste na identificação dos termos nos títulos dos artigos, isto é, no reconhecimento dos termos que se levrão em conta no processo de classificação. Isso se faz mediante a análise da sequência de caracteres do documento. Verficam-se termos incorretos e se remove a sequência de caracteres inválidos (caracteres de formatação de textos, por exemplo).

Posteriormente, excluem-se as *stopwords*. Nessa etapa, algumas palavras não são desconsideradas no processo de indexação, as preposições exemplificam esses tipos de palavras, pois são termos que conectam as ideias e as palavras. Então, calcula-se a "frequência relativa" de cada termo presente no texto (títulos dos artigos) de cada autor, o que consiste em dividir o número de vezes que um termo aparece nos títulos dos artigos de um autor pelo número total de termos relevantes que aparecem nos títulos.

Finalmente, para determinar a afinidade entre um autor e os assuntos (conceitos) presentes na ontologia, faz-se a multiplicação dos pesos dos termos presentes nos dois vetores (o vetor gerado a partir dos títulos dos artigos de um autor e cada vetor de assuntos presentes na ontologia), sendo que a soma desses produtos, limitada a um (1), representa o grau de similaridade existente entre o texto e o assunto.

Vários assuntos podem ser identificados como relacionados a um texto e, para cada um dos assuntos pode associar-se um grau (valor numérico) distinto que indica qual a probabilidade do documento conter o assunto.

Em um teste preliminar já apresentado em Ribeiro Jr. et al. (2005), a técnica utilizada obteve 91,66 % de acerto na classificação de textos de resumos retirados de artigos e 60,97 % de acerto na classificação de frases de resumos desses mesmos textos. Nesses testes, foram utilizados 15 artigos retirados de *CiteSeer-Directory* (2005), sendo consideradas certas as respostas nas quais o conceito identificado equivalia a entrada do diretório presente em *CiteSeer-Directoty* (2005).

A EXTRAÇÃO DE DADOS DA BDBCOMP

Partindo do princípio de que é possível identificar especialistas a partir de repositórios como bibliotecas digitais, optou-se por utilizar a BDBComp – Biblioteca Digital Brasileira de Computação (BDBComp, 2006). No momento da realização do experimento, a BDBComp possuía cerca de 4.580 trabalhos cadastrados.

Da BDBComp foram extraídas informações sobre os autores e os trabalhos publicados. Essas informações consistem, basicamente, na identificação

do autor (nome e código) e dos dados que determinam o trabalho, como título, autores e evento. Foi possível realizar a seleção de 1780 títulos de artigos.

Para obtenção das informações, elaborou-se um programa escrito em PHP sem interface gráfica que acessa o *site* da BDBComp e passa alguns argumentos iniciais de pesquisa para o *site* da BDBComp (o programa faz o processamento do retorno da consulta). Nesse caso, o retorno consiste em uma página HTML com os artigos de um autor que estão cadastrados na BDBComp. Dessas páginas foram extraídas algumas informações (nomes de autores, títulos de artigos e eventos em que os artigos foram publicados). O resultado desse processo armazena-se em uma base construída com MySQL, sendo, posteriormente, utilizado nos experimentos.

EXPERIMENTOS

Para a realização dos experimentos utilizou-se a ontologia e a técnica de *Text Mining* descritas na seção três. O texto utilizado como entrada para função de similaridade representou os títulos dos artigos de cada autor. Assim, para cada autor, aplicou-se a função de similaridade e o retorno indicou o grau de afinidade do autor em relação ao assunto presente na ontologia. Um autor, por exemplo, poderia possuir um grau de afinidade de 0.003190 em relação ao assunto/conceito Banco de Dados e 0.002445 em relação ao assunto/conceito Engenharia de *Software*. A partir disso, pode-se dizer que esse autor possui uma maior afinidade com a área de banco de dados do que com a área de engenharia de *software*.

Quando o processamento terminar, tornar-se possível descobrir quais os maiores especialistas em determinadas áreas. Logicamente, é preciso considerar que o exame limita-se ao conteúdo existente na BDBComp, o que pode, em algumas situações, fazer com que o resultado não seja real, desse modo, alguns especialistas podem não constar na análise ou ficar em uma posição inferior dentro do *ranking*. Contudo, espera-se que sejam identificadas pessoas que tenha um alto grau de afinidade em relação à determinada área do conhecimento, sem que sejam necessariamente os maiores especialistas da área. Nas considerações finais, essa discussão aprofundar-se-á.

Foram escolhidas quatro áreas para os experimentos: banco de dados, engenharia de *software*, inteligência artificial e redes de computadores. Para cada uma das áreas foram elencados cinco especialistas, ou seja, aqueles para os quais foi calculado um grau de afinidade maior em relação a cada uma das áreas.

Identificados esses especialistas, acessou-se o sistema Lattes do CNPq, usando o nome dos estudiosos mais selecionados, de modo a verificar se a identificação da especialidade feita pela ferramenta estava correta. Esse procedimento não foi automatizado, ou seja, não se desenvolveu nenhuma

ferramenta específica para realizá-lo. A identificação torna-se correta se no currículo Lattes de um autor presente na BDBComp, identificado como especialista em engenharia de software, por exemplo, estiver presente no item "Áreas de Atuação" referência à área de engenharia de *software*. Quando não se identificar essa informação, opta-se por tentar verificar se o especialista identificado possui uma *home page* contendo fortes indicativos de sua atuação na área indicada como sendo sua especialidade. Isso torna-se necessário em função de que uma pessoa pode, eventualmente, não preencher ou preencher de forma incompleta o campo "Áreas de Atuação" no currículo Lattes. As *home pages* dos especialistas localizam-se mediante pesquisas feitas no *Google*, pois seus nomes foram utilizados como argumentos para a pesquisa.

Para exemplificar o procedimento, identificou-se um autor que tinha um forte grau de afinidade com a área de Rede de Computadores, porém no item "Áreas de Atuação" do seu currículo Lattes essa especificação não constava. Já em sua *home page* explicitava-se esse interesse, uma vez que o autor, além de estudioso, ministrava disciplinas de Rede de Computadores.

Conforme mostra a tabela 1, para a área de Banco de Dados a técnica atingiu 100% de acerto, ou seja, os cinco especialistas encontrados pela técnica tinham, em seus Currículos Lattes ou em suas *home pages,* fortes indicativos de sua *expertise* área Banco de Dados. Esses resultados serão discutidos na seguinte seção.

Tabela 1. Resultado dos Experimentos.

Área/Assunto	Número de acertos	Percentual de acertos
Banco de Dados	5	100%
Engenharia de *Software*	5	100%
Inteligência Artificial	1	20%
Redes de Computadores	3	60%
GERAL	14	70%

CONSIDERAÇÕES FINAIS

Este trabalho demonstrou o resultado da aplicação de técnicas de *Text Mining* sobre o conteúdo de uma biblioteca digital, com o objetivo de identifi-

car especialistas em determinadas áreas do conhecimento. O método apresentou resultados satisfatórios, mas, aparentemente, a ontologia utilizada precisa rever alguns conceitos, tendo em vista os resultados obtidos. Existem, a princípio, correções a serem feitas nos termos e nos pesos relacionados à área de inteligência artificial. Essa preocupação no processo de construção da ontologia, tornou-se alvo de outros trabalhos realizados por alguns dos autores deste trabalho (LOH et al., 2006).

Conforme já esperado, constatou-se que a identificação de especialistas, a partir de uma biblioteca digital, pode distorcer alguns resultados. Desse modo, nem sempre os maiores especialistas de algumas áreas serão identificados, isso somente acontecerá se a biblioteca contiver um número considerável de publicações dos autores. No caso da BDBComp os trabalhos estão basicamente limitados àqueles publicados em eventos da SBC – Sociedade Brasileira da Computação, o que não permite com que se faça, por exemplo, a identificação (com uma maior probabilidade de acerto) dos maiores especialistas brasileiros em determinada área. Logicamente, os resultados se aprimorão à medida que um maior volume de informações adapta-se à base de dados.

Ainda, deve-se considerar que nem todos os trabalhos de um autor estão na BDBComp e que usar apenas o título dos artigos pode não ser suficiente. Sendo assim, para determinar quem tem maior conhecimento em determinada área deve-se, necessariamente, considerar a relevância do evento no qual um artigo foi apresentado, já que quanto maior for a relevância do evento mais rígida será a avaliação do artigo.

Embora que os resultados obtidos sejam preliminares, evidenciou-se a existência da possibilidade de descobrir quando alguém tem forte ou pelo menos considerável afinidade em relação a determinadas áreas pelo uso da técnica descrita. Nesse sentido, convém ressaltar que foi possível constatar que os donos dos currículos utilizados nos experimentos omitiram no item "Áreas de Atuação" do seu currículo Lattes informações sobre a sua afinidade com determinadas áreas. Esses dois casos demonstram, claramente, o ganho que pode ser obtido pela adoção das técnicas nesse estudo utilizadas. Caso se executasse uma consulta sobre os dados do currículo Lattes, usando XPath, XQuery ou SQL (no caso de o currículo estar armazenado em uma base de dados e não apenas em um documento XML), a fim de se encontrar pessoas com afinidade à área de Banco de Dados ou Redes de Computadores, buscar-se-ia essa informação apenas no item "Áreas de Atuação", o qual seria, talvez, o caminho mais natural. Porém, nesses caso os especialistas descritos não seriam identificados.

AGRADECIMENTOS

Este trabalho é parcialmente financiado pelo CNPq (Project DIGITEX - Editoração, Indexação e Busca em Bibliotecas, número 550845/2005-4), e pela FAPERGS (Projeto Rec-Semântica - Plataforma de Recomendação e Consulta na *Web* Semântica Evolutiva, número 0408933).

REFERÊNCIAS

BDBComp. **Building a Digital Library for the Brazilian Computer Science Community**. Disponível em: http://www.lbd.dcc.ufmg.br/bdbcomp/. Acesso em: Dez., 2006.

BECERRA-FERNANDEZ, I. **Facilitating the online search of experts at Nasa using expert seeker people-finder**, In: PAKM, v.34, 2000.

CITESEER-DIRECTORY. **Computer Science Directory – CiteSeer.** 2005. Disponível em: http://citeseer.ist.psu.edu/directory.html.

DAVENPORT, T. H; PRUZAC, L. **Working knowledge – How organizations managewhat they know**. Harvard: Harvard Business School Press, 1997.

D'AMORE, R. **Expertise tracking**, In: PROCEEDINGS OF 2005, INTERNATIONAL CONFERENCE ON INTELLIGENCE ANALYSIS, McLean, VA, 2005.

KRULWICH, B; BURKEY, C. **The contactfinder agent:** Answering bulletin board questions with referrals. In: PROCEEDINGS OF THE 1996 NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI-96), v.1, Portland, OR, p. 10-15, 1996.

LEWIS, D. D. **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, 1998.

LICHTNOW, D. et al. **Análise Automática de Textos em uma ComunidadeVirtual**, In: XXXI SEMISH - XXIV CONGRESSO DA SBC, Salvador, p. 202-216, 2004.

LOH, S. et al. **Constructing domain ontologies for indexing texts and creating user's profiles.** In: 1ST WORKSHOP ON ONTOLOGIES AND METAMODELING IN SOFTWARE AND DATA ENGINEERING - SBBD 2006, Florianópolis. p. 72-82, 2006.

LOH, D. et al. **Investigação sobre a Identificação de Assuntos em Mensagens de Chat**, In: TIL - XXIV CONGRESSO DA SBC, Salvador, p. 187-195, 2004.

LOH, S; WIVES, L. K; OLIVEIRA, J. P. M. **Concept-based Knowledge. Discovery in Texts Extracted from the Web,** ACM SIGKDD Explorations 2 (1), p. 29-39, 2000.

RIBEIRO JR. L. C. Identificação de áreas de interesse a partir da extração de informações de currículos lattes/xml. In: I ESCOLA REGIONAL DE BANCO DE DADOS, Porto Alegre. **Anais**... Porto Alegre : UFRGS, p. 67-72, 2005.

ROCCHIO, J. J. **Document Retrieval Systems - Optimization and Evaluation**. Ph.D. Thesis, Harvard Computation Laboratory, Harvard University, Report ISR-10 to National Science Foundation, 1996.

SALTON, G.; MCGILL, M. J. **Introduction to modern information retrieval**. New York; McGraw-Hill, 1983.

SOWA, J. F. **Building, Sharing, and Merging Ontologies**, AAAI Press / MIT press, p. 3-41, 2002.

STEETER, L. A; LOCHBAUM, K. E. **Who knows:** A system based on automatic representation of semantic structure. In: RIAO'88, Cambridge, MA, p. 380-388, 1988.