SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL
11 E 12 de Julho
Universidade Federal Fluminense • Departamento de Ciência da Informação
Niterói • Rio de Janeiro • Brasil

Esta comunicação está sendo submetida sob o
Tema 3 – Aplicações com enfoque em Ontologias

EXPLORANDO RELAÇÕES SEMÂNTICAS PARA EXPANSÃO DE CONSULTAS EM ACERVOS ESPECÍFICOS DE DOMÍNIO

EXPLORING SEMANTIC RELATIONS TO QUERY EXPANSION IN DOMAIN SPECIFIC DATASETS

André Bechara Elias (PPGI/UFRJ, andrebechara@yahoo.com.br)
Vanessa Braganholo (PPGI/UFRJ, braganholo@dcc.ufrj.br)
Maria Luiza Machado Campos (PPGI/UFRJ, mluiza@nce.ufrj.br)

Resumo: Com a popularização da Web, os mecanismos de recuperação de informação se tornaram essenciais no cotidiano das pessoas. Apesar da rápida evolução destes mecanismos nos últimos anos, ainda existem muitas oportunidades de pesquisa nesta área. As ontologias abrem uma nova oportunidade em pesquisas de recuperação de informação não-estruturada. O conhecimento mapeado dentro de uma ontologia pode ser utilizado em diversas etapas do processo de recuperação de informação. Este artigo propõe uma técnica de expansão de consultas baseada em ontologias de domínio. Além disso, é avaliada de que forma as relações existentes nas ontologias afetam a eficiência da técnica.

Palavras-chave: Recuperação de Informação. Ontologia. Expansão de Consulta.

Abstract: *With the popularization of the Web, mechanisms for information retrieval have become essential in the daily lives of people. Despite the rapid evolution of these mechanisms in recent years, there are still many opportunities for research in this area. Ontologies have opened new opportunities in non-structured information retrieval research. Knowledge mapped within an ontology can be used in various stages of the process of information retrieval. This paper proposes a technique for query expansion based on domain ontologies. We also investigate how the existing semantic relations in ontologies affect the efficiency of the technique.*

Keywords: *Information Retrieval. Ontology. Query Expansion.*

EXPLORANDO RELAÇÕES SEMÂNTICAS PARA EXPANSÃO DE CONSULTAS EM ACERVOS ESPECÍFICOS DE DOMÍNIO

Resumo: Com a popularização da Web, os mecanismos de recuperação de informação se tornaram essenciais no cotidiano das pessoas. Apesar da rápida evolução destes mecanismos nos últimos anos, ainda existem muitas oportunidades de pesquisa nesta área. As ontologias abrem uma nova oportunidade em pesquisas de recuperação de informação não-estruturada. O conhecimento mapeado dentro de uma ontologia pode ser utilizado em diversas etapas do processo de recuperação de informação. Este artigo propõe uma técnica de expansão de consultas baseada em ontologias de domínio. Além disso, é avaliada de que forma as relações existentes nas ontologias afetam a eficiência da técnica.

Palavras-chave: Recuperação de Informação. Ontologia. Expansão de Consulta.

Abstract: *With the popularization of the Web, mechanisms for information retrieval have become essential in the daily lives of people. Despite the rapid evolution of these mechanisms in recent years, there are still many opportunities for research in this area. Ontologies have opened new opportunities in non-structured information retrieval research. Knowledge mapped within an ontology can be used in various stages of the process of information retrieval. This paper proposes a technique for query expansion based on domain ontologies. We alse investigate how the existing semantic relations in ontologies affect the efficiency of the technique.*

Keywords: *Information Retrieval. Ontology. Query Expansion.*

1. Introdução

Ontologias têm sido usadas como apoio para acesso a acervos específicos de domínio. Várias abordagens utilizam ontologias ou taxonomias para a indexação de acervos, de forma a facilitar a navegação dentro do acervo. Exemplos podem ser encontrados em bases de dados de grandes empresas, como a Petrobrás, em bases de legislação, em sites na internet, entre outros.

No entanto, estes usos têm se restringido à navegação. Isso significa que, para encontrar o que deseja, o usuário tem que navegar pelas categorias da taxonomia (ou classes da ontologia), utilizando os relacionamentos hierárquicos até encontrar os documentos que deseja. Dependendo da profundidade da hierarquia e do número de conceitos que ela disponibiliza, a tarefa de navegação pode se tornar cansativa e pouco eficiente.

Por outro lado, máquinas de busca têm sido utilizadas para facilitar a tarefa de busca por informações. Tais máquinas se baseiam nas teorias da área de Recuperação de Informação (Baeza-Yates e Ribeiro-Neto, 1999) e são capazes de realizar consultas por palavras-chave no acervo de documentos. Os resultados são exibidos ao usuário por ordem de relevância, ou seja, os documentos que a máquina de busca *julga* serem mais importantes são colocados no topo da lista de resultado. Este julgamento de relevância leva em conta fatores como o número de vezes que um termo aparece no acervo, entre

outros, mas não leva em consideração a *semântica* dos termos presentes nos documentos. Isso faz com que, em grande parte das vezes, resultados importantes não sejam retornados para o usuário, simplesmente porque aquele documento não continha o termo utilizado na busca, mas sim um sinônimo ou um termo mais genérico. Por exemplo, se um usuário busca por "livro", documentos que contenham a palavra "romance", mas não "livro", não serão retornados para o usuário. A máquina de busca não "sabe" que "romance" é um tipo de "livro".

Uma das formas de amenizar este problema consiste em aplicar técnicas de expansão de consultas. No exemplo anterior, a consulta poderia ser expandida para consultar "livro" e "romance" ao invés de apenas "livro". Isso faz com que novos documentos sejam retornados, aumentando a cobertura do sistema (fração de documentos relevantes que foram recuperados por uma busca).

Vários trabalhos na literatura mostraram que a técnica de expansão de consultas pode ter efeito negativo sobre o sistema, diminuindo a precisão do sistema (a fração dos documentos recuperados que são relevantes). Vários parâmetros influenciam nos resultados da técnica: quais características considerar para a expansão das consultas?; quantos termos devem ser utilizados na expansão?; de onde retirar os termos que serão expandidos? Nós acreditamos que as relações semânticas presentes em uma ontologia (subtipo_de, instância_de, sinonimo_de, supertipo_de) podem ajudar na expansão de consultas, aumentando a cobertura e a precisão do sistema. Neste trabalho, propomos explorar os vários tipos de relações semânticas existentes em uma ontologia e estudar quais deles melhoram a qualidade dos resultados da busca quando utilizados para expandir consultas de uma base específica de domínio.

O restante deste trabalho está organizado como segue. A seção 2 discute os trabalhos relacionados na literatura. Na seção 3 mostraremos como será feita a análise das relações presentes na ontologia. Na seção 4 apresentamos considerações finais e as próximas etapas deste trabalho.

2. Trabalhos Relacionados

Experimentos de recuperação de informação que exploram relações semânticas entre termos têm sido realizados há décadas. Um dos primeiros experimentos publicados (Salton e Lesk, 1968) já explorava estas relações em pequenos acervos específicos de domínio com o auxilio de um tesauro. Salton realizou vários experimentos utilizando a mesma metodologia. Destes experimentos, o experimento

Cranfield-I (CRAN-1) é o de maior relevância para acervos específicos de domínio. Uma das etapas do experimento CRAN-1 analisou de que forma as relações de sinonímia impactariam a eficiência da recuperação de informação. Os resultados mostraram que estas relações produzem melhoras significativas na recuperação de informação. Salton salienta ainda que a qualidade do tesauro tem alta correlação com este ganho de desempenho. Apesar dos resultados indicarem claramente que a técnica é eficiente, algumas considerações devem ser feitas. Inicialmente, é importante lembrar que a coleção de teste CRAN-1 (200 documentos) é muito pequena quando comparada à maioria dos acervos reais específicos de domínio. Do ponto de vista das consultas, a média de 17 termos que elas utilizavam está muito distante da média de 2,2 termos inseridos por usuários de sistemas de recuperação de informação em suas consultas (Spink *et al.*, 2001). Por fim, o tesauro utilizado foi construído por poucos especialistas de domínio no momento do experimento. Desta forma, as relações entre os termos do domínio não tinham o grau de formalismo adequado para o experimento.

Um dos primeiros trabalhos a explorar de forma profunda o impacto das relações entre os termos na técnica de expansão de consultas foi o experimento de Fox (1980). Os resultados obtidos mostraram que em geral a expansão de consultas baseada em relações léxicas melhorou a precisão média do mecanismo de recuperação em torno de 20%. Além disso, o experimento mostrou que as expansões baseadas em antônimos acabaram por degradar o desempenho do mecanismo de recuperação. No entanto, apesar de servir com um excelente framework de pesquisa para esses tipos de técnica, o acervo utilizado também foi muito pequeno. Além disso, as relações foram testadas em grupos, deixando em aberto uma análise mais profunda de cada relação léxica individualmente.

Experimentos mais recentes utilizaram bases maiores, mais próximas da realidade. O experimento de Voorhees (1994) usou um tesauro livre de domínio (o WordNet) como apoio para as expansões dos termos das consultas e a base TREC como coleção de testes. Esta base consiste em um acervo livre de domínio totalizando 742000 documentos, os quais foram extraídos de jornais e revistas técnicas. Foram utilizadas as relações léxicas de sinonímia, hiperonímia, hiponímia e relacionado_a presentes na WordNet para expansão. Além disso, foram testadas várias combinações de pesos para os termos adicionados pelo sistema. O experimento foi realizado sobre o sistema de recuperação de informação SMART, que é baseado no modelo vetorial clássico (Baeza-Yates e Ribeiro-Neto, 1999). De forma a selecionar os termos corretos para a expansão,

os conjuntos de sinônimos foram escolhidos manualmente (por um ser humano) levando-se em consideração todo o contexto da consulta. Desta forma, os resultados reportados representam um limite superior do desempenho esperado de um mecanismo de expansão de consultas automático que utiliza esta estratégia de expansão. Neste experimento, Voorhees conseguiu uma melhora de 35% na precisão média de 11 pontos na sua execução de menor número de termos da consulta. Nesta execução as médias de termos de consulta foram 11,02. Os pesos utilizados nesta execução foram 1 para os termos informados na consulta e 0,5 para os termos expandidos. Este resultado mostrou que o vetor de consulta deve privilegiar os termos inseridos pelo usuário em relação aos termos expandidos automaticamente.

3. Análise de Relações

Para analisar o impacto das relações semânticas na expansão de consultas, estamos realizando um experimento prático usando como base de dados a base TREC Genômica. Esta base compreende todos os documentos publicados no Medline entre 1994 até 2004, totalizando 4,6 Milhões de documentos. O Medline é o principal acervo utilizado por pesquisadores da área Biomédica. Para obter as relações semânticas deste domínio, estamos utilizando a GeneOntology.

Para uma melhor análise dos resultados, eliminamos variáveis que possam distorcer o resultado da técnica, como por exemplo, expansão automática dos termos. Por causa das relações de polissemia, a correta expansão dos termos envolve análise de contexto, e portanto, torná-la automática por meios computacionais envolve um estudo minucioso e é por conseguinte tema para trabalhos futuros. Porém, acreditamos que os resultados obtidos neste trabalho poderão auxiliar a construção de um algoritmo de expansão automática baseado em ontologias.

Nossos experimentos estão sendo conduzidos da seguinte forma. Inicialmente indexamos a coleção de testes TREC Genômica. Posteriormente, estamos expandindo manualmente as 50 consultas da coleção de testes utilizando a Gene Ontology como apoio para a expansão. Para cada consulta foi verificada a existência das seguintes relações: subtipo_de, instância_de, sinonimo_de, supertipo_de. Desta forma, foram geradas 200 consultas expandidas, cada uma utilizando um tipo de relação semântica. Note que em cada consulta, não há mistura de tipos de relação. Esta preocupação é necessária para que se possa avaliar os resultados obtidos por cada tipo de relação

separadamente. As consultas resultantes da expansão serão submetidas a um mecanismo de recuperação de informação padrão para colher medidas de precisão e cobertura. Estes resultados serão comparados aos resultados obtidos pelo mesmo mecanismo sem utilizar a técnica de expansão.

Finalmente serão avaliados os casos de sucessos e de falhas, por consulta e por relação semântica, para apontar as possíveis causas da degradação ou melhora dos resultados. Acreditamos que ao final do experimento teremos gerado uma base teórica que possibilitará a compreensão de como e quando expandir os termos de consulta específica de domínio inseridos pelo usuário.

4. Considerações Finais

Este artigo apresenta uma proposta de utilização de ontologias de domínio para apoiar as técnicas de expansão automática de consultas. Para colher indícios da eficiência desta técnica está sendo realizado um experimento sobre a base TREC Genômica e as relações semânticas da Gene Ontology. Estamos finalizando a etapa de expansão de consultas e nos comprometemos a apresentar os primeiros resultados por ocasião da apresentação do artigo no evento.

Referências

Baeza-Yates, R.; Ribeito-Neto, B. *Modern Information Retrieval*. New York: ACM Press, 1999.

Fox, E. *Lexical relations enhancing effectivness of information retrieval systems*. SIGIR Forum, New York, v.15, n.3, p.5-36, 1980.

Salton, G.; Lesk, M. *Computer evaluation of indexing and text processing*. New Jersey, Prentice-Hall, 1971.

Spink, A.; Wolfram, D.; Jansen, M.B.J.; Saracevic, T. *Searching the Web: The Public and Their Queries*. Journal of American Society for Information Science and Technology, [s.i.], n.52(3), p.226-234, 2001

Voorhees, E. *Query expansion using lexical-semantic relations*. In: ACM SIGIR conference on research and development in information retrieval, Procedings, Dublin:17, p.61–69, 1994