Expansão Automática de Consultas utilizando Ontologias

Automatic Query Expansion with Ontologies

Fabrício Jailson Barth, Antonio Pedro Timoszczuk

Diretoria de Inovações e Soluções Tecnológicas
Fundação Atech Tecnologias Críticas (http://www.atech.br)
Rua do Rocio, 313, 11º andar, Vila Olímpia, São Paulo – SP

{fbarth, antoniop} @atech.br

Resumo: Este trabalho descreve um mecanismo de recuperação de informações que utiliza uma ontologia de domínio para gerar consultas contextualizadas. A critério do usuário, a consulta original pode ser refinada mediante duas operações: focalização e generalização. Os resultados obtidos mostram que as operações que fazem uso da ontologia possuem, na maioria dos casos, uma *medida F* superior à da busca sem contextualização. Além disso, observou-se que a operação generalizada obtém uma consulta mais abstrata enquanto que a operação focalizada obtém uma consulta mais específica, permitindo ao usuário navegar a ontologia tendo em vista o ajuste de sua consulta ao nível de abstração desejada.

Palavras-chave: Ontologias, Sistemas de Recuperação de Informação, Expansão de Consultas.

Abstract: *This paper describes a retrieval information algorithm that uses a domain ontology to turn out contextualize queries. Queries can be contextualized by two operations: focalization and generalization. Experiment results shows that contextualized operations have a F measure bigger than operations that do not use a domain ontology. Otherwise, we can seen that operations have different levels of abstraction, given to the user the possibility to adjust the query abstract level.*

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

Introdução

Recuperação de Informação (RI) é a parte da Ciência da Computação que estuda a recuperação de informações de uma coleção de documentos. Os documentos recuperados estão destinados a satisfazer a necessidade de informação de um usuário geralmente expressa em linguagem natural [1, 5]. Os Sistemas de Recuperação de Informações disponíveis atualmente permitem que o usuário obtenha informações com relativa rapidez e facilidade. No entanto, alguns problemas ainda devem ser solucionados para que sistemas mais eficientes venham a produzir informações mais relevantes para o usuário [3]. Alguns estudos relatam que os usuários realizam consultas com poucos termos, em média dois, criando consultas ambíguas e contribuindo para: *(i)* um desvio significativo entre a ordenação dos resultados (*ranking*) produzida pelos Sistemas de Recuperação de Informação e aquela que o usuário desejaria, e; (*ii*) o retorno de centenas de resultados pouco relevantes para o usuário [4].

Alguns Sistemas de Recuperação de Informação expandem a consulta do usuário, de maneira semi-automática ou automática, visando diminuir os problemas de ambigüidade provocados por consultas com poucos termos. Diversos métodos para expansão de consultas têm sido propostos. Estes métodos são divididos por três categorias: métodos que utilizam a interação com o usuário para expandir a consulta, métodos que utilizam documentos recuperados anteriormente pelo usuário e métodos que utilizam uma coleção de documentos pré-determinada, local ou não [4].

O objetivo deste trabalho é verificar a aplicabilidade do uso de Ontologias para a expansão automática de consultas em Sistemas de Recuperação de Informação. Na próxima seção é apresentado o algoritmo para expansão de consultas, na seção seguinte são descritos os resultados alcançados e na última seção são apresentadas as considerações finais do trabalho.

Material e métodos

Neste trabalho, para aumentar a precisão e o índice de recuperação de documentos relevantes é utilizada uma ontologia de domínio. Ao permitir a expansão da consulta do usuário a partir de termos adicionais próprios do domínio, espera-se que a ontologia viabilize a recuperação de documentos que seriam ignorados pela consulta original.

No contexto da engenharia do conhecimento, uma ontologia é especificada sob a forma de um vocabulário que representa os conceitos do domínio [6]. Um exemplo simples, que corresponde à ontologia desenvolvida neste projeto, é o da hierarquia de tipos, onde são especificados classes (conceitos) e seus relacionamentos com superclasses e subclasses. Esse tipo de ontologia resulta, portanto, da decomposição do domínio em conceitos que se relacionam com outros mais genéricos e mais específicos. Todos os conceitos podem ter sinônimos associados. Os sinônimos permitem a cobertura exaustiva do vocabulário do domínio, enquanto que a hierarquia de tipos permite a navegação de conceitos mais específicos para mais genéricos, e vice-versa.

A consulta submetida pelo usuário é contextualizada da seguinte maneira: cada termo que compõe a consulta é pré-processado tendo em vista a remoção de acentos, de maiúsculas e a redução ao singular. Em seguida, o termo é confrontado com aqueles definidos na ontologia e seus sinônimos. Se o termo ou um de seus sinônimos for encontrado na ontologia, todos são conectados com o operador *OR*, compondo uma

expressão que vai substituir o termo original na consulta. Caso o termo não esteja na ontologia, ele é mantido sem alterações na consulta. Todos os termos, expandidos ou não, são conectados com o operador *AND* para formar a consulta contextualizada. Termos compostos podem fazer parte tanto da consulta como da ontologia. Tais termos, que devem ser especificados entre aspas pelo usuário, não são reduzidos ao singular.

A critério do usuário, a consulta original ou contextualizada pode ser refinada mediante duas operações: focalização e generalização [2]. Na generalização, os termos da consulta e seus sinônimos são substituídos pelos termos correspondentes à(s) superclasse(s) e seus sinônimos. Na focalização, são acrescentados aos termos da consulta e seus sinônimos os termos correspondentes à(s) subclasses(s) e seus sinônimos. O objetivo da primeira operação é obter uma consulta mais abstrata (composta de termos mais genéricos na hierarquia); o da segunda, obter uma consulta mais específica (composta de termos mais específicos na hierarquia). Isto permite ao usuário navegar a ontologia tendo em vista o ajuste de sua consulta ao nível de abstração desejado.

Resultados

A avaliação do algoritmo de recuperação de informação contou com a colaboração de dois especialistas do domínio onde o Sistema de Recuperação de Informação foi utilizado. Cada especialista envolvido na avaliação formulou cinco consultas e identificou, para cada consulta formulada, os documentos que considerava relevantes entre os 100 documentos filtrados aleatoriamente antes do início da avaliação.

Para cada consulta, foram aplicadas as três operações implementadas (contextualizar, focalizar e generalizar) e uma outra operação onde a consulta não foi alterada em função da ontologia, chamada de consulta simples. Os resultados obtidos mostram que as operações que fazem uso da ontologia (contextualizar, focalizar e generalizar) possuem, na maioria dos casos, uma *medida F*[1] superior à da busca simples. A média da *medida F*

[1] A *medida F* é uma forma de definir uma média harmônica entre os índices de precisão e recuperação [4]: Medida F = (2 x precisão x recuperação) / (precisão + recuperação). Precisão é a proporção dos documentos recuperados que são relevantes para uma dada consulta em relação ao total de documentos recuperados. O índice de recuperação é a razão entre o número de documentos recuperados que são relevantes para uma consulta e o total dos documentos na coleção que são considerados relevantes para a consulta [1].

da busca simples é *0.41*, enquanto que as médias da *medida F* das buscas contextualizadas, focalizadas e generalizadas são: *0.48*, *0.54* e *0.52*, respectivamente.

Levando-se em consideração que o objetivo da operação generalizada é obter uma consulta mais abstrata e o da operação focalizada é obter uma consulta mais específica. Em termos de medida de precisão e recuperação, isto significa dizer que: a consulta generalizada deve ter um índice de precisão mais baixo e um índice de recuperação mais alto, e; a consulta focalizada teve ter um índice de precisão mais alto e um índice de recuperação mais baixo.

Com a execução dos experimentos foi possível constatar que: a média da precisão da consulta focalizada (*0.54*) é maior que a média da precisão da consulta generalizada (*0.46*), e; a média do índice de recuperação da consulta focalizada (*0.74*) é menor que a média do índice de recuperação da consulta generalizada (*0.88*).

Discussão e Conclusão

Este trabalho apresentou um mecanismo de recuperação de informações que utiliza uma ontologia de domínio para gerar consultas contextualizadas. A análise dos resultados obtidos demonstrou que as operações que fazem uso da ontologia (contextualizar, focalizar e generalizar) possuem um desempenho superior quando comparadas a um processo de recuperação de informação convencional. Além disso, observou-se que a operação generalizada obtém uma consulta mais abstrata enquanto que a operação focalizada obtém uma consulta mais específica, permitindo ao usuário navegar a ontologia tendo em vista o ajuste de sua consulta ao nível de abstração desejada.

As principais dificuldades encontradas durante a execução do projeto relacionaram-se ao ajuste da ontologia, tendo em vista a finalidade daquela descrição do domínio: estruturar adequadamente os termos do domínio tal como empregados nas fontes de informação utilizadas.

A principal restrição deste trabalho está relacionada com a coleção de referência utilizada para a validação da proposta. Os pontos negativos desta coleção são: (i) é uma coleção com poucos documentos e poucas consultas, e; (ii) é uma coleção criada apenas para esta avaliação. Como trabalhos futuros pretende-se avaliar esta mesma proposta utilizando coleções de referência amplamente conhecidas na área de Recuperação de Informação.

Referências

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. ACM Press, 1999.

[2] Dario Bonino, Fulvio Corno, Laura Farinetti, and Alessio Bosca. Ontology driven semantic search. *WSEAS Transaction on Information Science and Application*, 1(6):1597–1605, December 2004.

[3] Czeslaw Danilowicz and Huy Cuong Nguyen. Using user profiles in intelligent information retrieval. In Mohand-Saïd Hacid, Zbigniew W. Rás, Djamel A. Zighed, and Yves Kodratoff, editors, *Foundations of Intelligent Systems*. 13th International Symposium, number LNAI 2366, pages 223–231, Lyon, France, June 2002. Springer-Verlag.

[4] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[5] Tefko Saracevic. Evaluation of evaluation in information retrieval. In: *SIGIR '95: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM Press, 1995. p. 138-146.

[6] Mike Uschold and Michael Gruninger. Ontologies: Principles, methods and applications. *The Knowledge Engineering Review*, 11(2):93–155, 1996.