Recuperação de informação usando ontologias geográficas relacionadas

Conference Paper · September 2009

2 authors, including:

Ivan Ricarte
University of Campinas
132 PUBLICATIONS **320** CITATIONS

SEE PROFILE

RECUPERAÇÃO DE INFORMAÇÃO UTILIZANDO ONTOLOGIAS GEOGRÁFICAS RELACIONADAS

MARIA ANGELICA DE ANDRADE LEITE[1]
IVAN LUIZ MARQUES RICARTE[2]

RESUMO: Com a crescente popularidade da *World Wide Web* mais pessoas têm acesso à informação cujo volume vem expandindo ao longo do tempo. O acesso a esta informação, de forma eficiente, é fundamental para que o conhecimento seja difundido ajudando os processos de tomada de decisão. Neste contexto a área de recuperação de informação ganhou um novo desafio visando buscar a informação, de uma forma mais inteligente, pelo significado da informação contida nos documentos. Uma forma de recuperar a informação, pelo seu significado, é pelo uso de uma base de conhecimento que modela os conceitos de um domínio e seus relacionamentos. Atualmente, ontologias têm sido utilizadas para modelar bases de conhecimento. Este trabalho apresenta um método de expansão de consulta que utiliza uma base de conhecimento composta de ontologias geográficas relacionadas para melhorar o processo de recuperação de informação. O método de expansão de consulta é testado com a máquina de busca do Apache Lucene mostrando uma melhoria na qualidade da informação recuperada.

PALAVRAS-CHAVE: recuperação de informação geográfica, ontologia, representação do conhecimento, expansão da consulta.

INFORMATION RETRIEVAL USING RELATED GEOGRAPHIC ONTOLOGIES

ABSTRACT: With the World Wide Web popularity growth, more people has access to information and this information volume is expanding over the time. The information access, in a efficient way, is fundamental to knowledge diffusion helping the decision making process. In this context, the information retrieval area has a new challenge intending to search information, in a more intelligent way, by the documents meaning. A way to retrieve information, by its meaning, is by using a knowledge base that encodes the domain concepts and their relationships. Nowadays, ontologies are being used to model knowledge bases. This work presents a query expansion method that uses a knowledge base comprised of related geographic ontologies to improve the information retrieval process. The query expansion method is tested with the Apache Lucene search engine showing an improvement in the retrieved information quality.

KEY-WORDS: geographic information retrieval, ontology, knowledge representation, query expansion.

1. INTRODUÇÃO

Um sistema de recuperação de informação armazena e indexa documentos de forma que quando os usuários expressam sua necessidade de informação em uma consulta, o sistema recupera os documentos relacionados associando um *score* a cada um. Quanto maior o *score* maior a relevância do documento (Baeza-Yates et al., 1999). Usualmente um sistema de recuperação de informação retorna um conjunto muito grande de documentos e os usuários

[1] Engenheira da Computação, Embrapa Informática Agropecuária, angelica@cnptia.embrapa.br
[2] Engenheiro Eletricista, Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas , ricarte@fee.unicamp.br

precisam empregar um tempo considerável até encontrar aqueles que sejam realmente relevantes. Além disto, os documentos são recuperados quando eles contêm as palavras-chave especificadas na consulta. Entretanto, este enfoque negligencia outros documentos, também relevantes para a consulta, mas que não contêm os termos especificados na mesma. Ao considerar um domínio de conhecimento específico este problema pode ser superado pela incorporação de uma base de conhecimento no processo de recuperação de informação. A base de conhecimento vai modelar os conceitos de um domínio e seus relacionamentos. Ao utilizar uma base de conhecimento o objetivo é utilizar o conhecimento expresso na base para fazer a expansão da consulta do usuário. A expansão de consulta consiste em adicionar novos termos semanticamente relacionados, com os termos presentes na consulta inicial, em função do conhecimento contido na base de conhecimento. A expectativa é melhorar a qualidade dos documentos recuperados trazendo mais documentos associados à consulta (melhoria da taxa de cobertura) e apresentando estes documentos numa ordem onde os documentos do topo da lista de documentos sejam os mais relevantes à consulta (melhoria da taxa de precisão) (Leite et al., 2008; Leite, 2009).

2. OBJETIVO

Neste artigo é mostrado um método de expansão da consulta no domínio geográfico considerando uma ontologia de divisão territorial e uma ontologia de clima. As máquinas de busca convencionais tratam os nomes de lugares da mesma forma que as outras palavras-chave e irão recuperar os documentos que contêm o nome especificado. Em alguns contextos uma região é caracterizada pelas características que ela possui e os documentos associados a estas características são interessantes serem recuperados mesmo que o seu nome não esteja presente. Por exemplo, a Região Nordeste, no Brasil, possui o clima semi-árido e em muitas situações a região é referenciada apenas como região semi-árida ao invés do seu nome geográfico. Assim, documentos contendo informação sobre o clima semi-árido são interessantes serem recuperados para uma consulta associada à Região Nordeste. Por um outro lado, a informação geográfica explícita pode não estar presente no documento como, por exemplo, a indicação de uma entidade geográfica mais geral ou mais específica é omitida embora ela esteja subentendida pelos usuários. O método de expansão de busca proposto é empregado juntamente com máquina de busca do Apache Lucene (Apache Project, 2009) numa coleção de documentos no domínio da agrometeorologia. Os resultados obtidos com o método de expansão da consulta são comparados com os resultados utilizando apenas as palavras-chave da consulta inicial. Os resultados mostram uma melhora na qualidade dos documentos recuperados. A expectativa é acrescentar, na base de conhecimento, uma ontologia que modele os recursos naturais, presentes no território brasileiro, permitindo uma recuperação de documentos mais inteligente neste domínio.

3. MÉTODO DE EXPANSÃO DA CONSULTA

O conhecimento é representado por ontologias relacionadas onde cada uma corresponde a um domínio distinto. Cada domínio é um conjunto de conceitos $D_k = \{c_{k1}, c_{k2}, ... c_{ky}\}$ onde $1 \leq k \leq K$, K é o número de domínios e $y = |D_k|$ é o número de conceitos em cada domínio. Estas ontologias estão relacionadas compondo a base de conhecimento. Os conceitos dentro das ontologias estão organizados como uma taxonomia e são relacionados por associação de especialização (S) e associação de generalização (G). Os conceitos pertencentes a ontologias distintas estão relacionados por associação positiva (P). A figura 1 ilustra o esquema de representação do conhecimento com uma amostra das ontologias geográficas e o relacionamento entre elas. O domínio D_1 corresponde à ontologia de divisão territorial e o domínio D_2 corresponde à ontologia de clima. As ontologias foram construídas a partir do

conhecimento existente no mapa que trata da distribuição de clima Köppen no território brasileiro (SISGA, 2009).

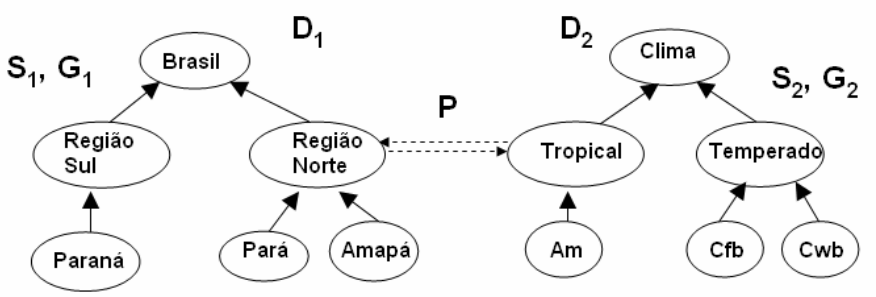

Figura 1: Representação do conhecimento.

O conjunto de documentos é dado por DOC = $\{d_1, d_2, d_3, ..., d_n\}$ onde $1 \leq n \leq N$ e N é o número de documentos da coleção. Os documentos são indexados pela máquina de busca Lucene. A cada conceito $c_{ky,}$ das ontologias, presente no documento, é associado um valor w_{ky}^n que é um número real onde $w_{ky}^n \geq 0$. O valor w_{ky}^n é calculado seguindo o esquema *tf-idf* (Manning et al., 1999). O valor w_{ky}^n indica o quanto o conceito c_{ky} representa o conteúdo do documento d_n. Quanto maior o valor w_{ky}^n maior é a associação entre o conteúdo do documento d_n e o conceito c_{ky} da ontologia.

As consultas são expressas com os conceitos dos domínios conectados por operadores lógicos como E ou OU. O método de expansão considera o conhecimento expresso pelas ontologias. Como os domínios estão representados por ontologias distintas então as consultas, q, também devem ser particionadas para considerar os conceitos de cada domínio separadamente. Cada partição será constituída por um vetor $q = (u_{k1}, u_{k2}, ..., u_{ky})$ onde $1 \leq k \leq K$ e $1 \leq y \leq | D_k |$. O valor $u_{ky} \in \{0, 1\}$ indica a presença (1) ou a ausência (0), do respectivo conceito $c_{ky} \in D_k$, na consulta do usuário. Dados os domínios D_1 = {Brasil, Região Sul, Paraná, Região Norte, Pará, Amapá} e D_2 = {Clima, Tropical, Am, Temperado, Cfb, Cwb} uma consulta expressa por q = Região Norte seria representada como q = (0 0 0 1 0 0).

Através da base de conhecimento é possível explorar as relações entre os conceitos dos domínios para expandir a consulta do usuário com novos conceitos que, apesar de não estarem presentes na consulta inicial, sejam relacionados a estes. Pela expansão da consulta os novos conceitos serão incorporados à mesma permitindo a recuperação de documentos que sejam semanticamente relacionados à consulta original, em função do conhecimento contido na base. A expansão ocorre em duas fases. Na primeira fase a associação positiva (P) é utilizada. A figura 2 ilustra a primeira fase da expansão da consulta considerando a base de conhecimento formada pelos domínios D_1 e D_2. Nesta base de conhecimento os conceitos Região Norte e Tropical estão associados pela associação positiva. Depois da primeira fase da expansão tem-se o valor da consulta expandida qent = (Região Norte ou Tropical).

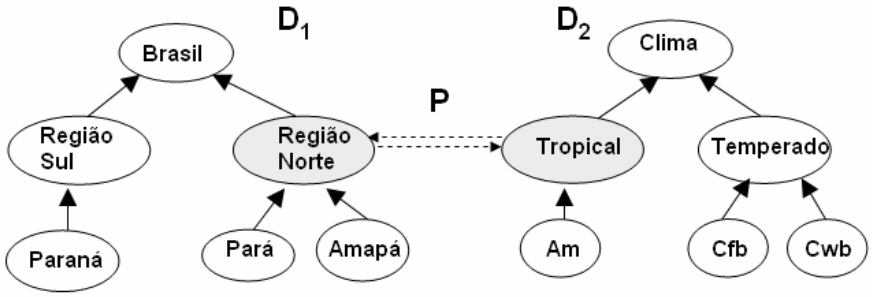

Figura 2: Primeira fase da expansão da consulta.

Uma vez que se tenha a consulta qent expandida entre os domínios é realizada a segunda etapa da expansão. Esta etapa visa realizar a expansão da consulta qent intradomínios, isto é, considerando as relações de especialização (S) e generalização (G) em cada ontologia. A figura 3 ilustra a segunda fase da expansão da consulta. Na ontologia que representa o domínio D_1 tem-se que o conceito Brasil é mais geral que o conceito Região Norte e os conceitos Pará e Amapá são mais específicos. Na ontologia que representa o domínio D_2 tem-se que o conceito Clima é mais geral que o conceito Tropical e o conceito Am é mais específico. Nesta fase da expansão, os conceitos mais específicos e mais gerais, que os conceitos presentes em qent, são adicionados à consulta. Depois da segunda fase da expansão tem-se a representação final da consulta expandida dada por qexp = (Brasil ou Região Norte ou Pará ou Amapá) ou (Clima ou Tropical ou Am). É possível associar pesos aos conceitos permitindo recuperar documentos que sejam mais associados a alguns conceitos do que a outros. Neste caso, por exemplo, pode-se associar um peso maior ao conceito Pará que ao conceito Brasil pois o conceito Pará, por ser mais específico, está mais fortemente relacionado com a Região Norte do que o conceito Brasil.

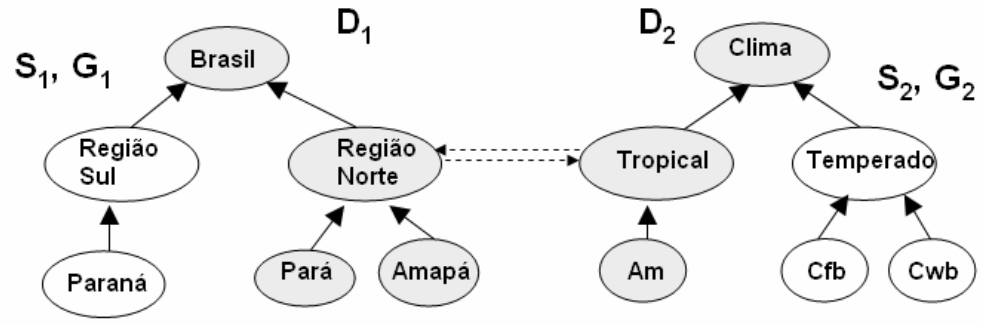

Figura 3: Segunda fase da expansão da consulta.

Depois que a consulta é expandida, a relevância dos documentos com relação à consulta é calculada. A função de relevância associa um *score* para cada documento dependendo do quão importante o documento é para a consulta. Os documentos são ordenados em ordem decrescente do seu *score* e são apresentados ao usuário.

4. RESULTADOS E DISCUSSÃO

A avaliação experimental foi realizada considerando uma amostra de 129 documentos selecionados da coleção de documentos do domínio da Agrometeorologia no Brasil, mantida pela Embrapa, e um conjunto de 83 consultas. O experimento utilizou o método de expansão da consulta considerando a máquina de busca Lucene do projeto Apache. O desempenho foi avaliado utilizando as medidas de precisão e cobertura. Dada uma consulta, a precisão é a fração do número de documentos relevantes recuperados pelo número total de documentos recuperados e a cobertura é a fração do número de documentos relevantes recuperados pelo número total de documentos relevantes na coleção. Os resultados foram expressos no gráfico de precisão *versus* cobertura da figura 4. Este gráfico faz uma comparação do desempenho da busca utilizando apenas as palavras-chave e utilizando o método de expansão da consulta proposto.

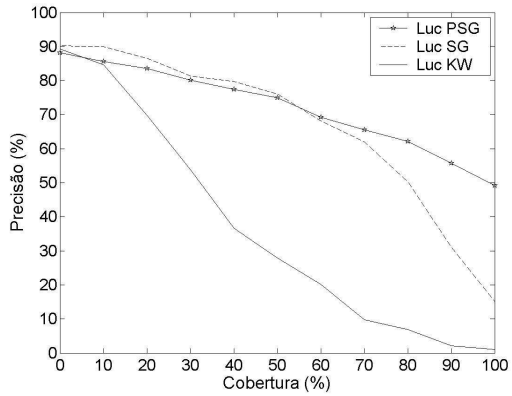

Figura 4: Gráfico de precisão *versus* cobertura.

No gráfico da figura 4, a curva LucKW representa as consultas compostas apenas pelas palavras-chave, a curva LucSG representa as consultas expandidas com os conceitos dentro das ontologias, considerando apenas as associações de especialização (S) e generalização (G), e a curva LucPSG representa a consulta expandida considerando os três tipos de associação da base de conhecimento. Neste gráfico pode-se observar que, quando o valor da cobertura é pequeno o valor da precisão é similar para os três tipos de consulta. À medida que o valor da cobertura aumenta o melhor valor de precisão é obtido quando os três tipos de associação (PSG), entre os conceitos das ontologias, são considerados. Este fato indica que o uso de uma base de conhecimento composta por ontologias relacionadas melhora o desempenho da recuperação de informação.

5. CONCLUSÕES

Este artigo mostrou um método de expansão de consulta utilizando uma base de conhecimento composta por ontologias geográficas relacionadas. Os resultados experimentais mostraram que houve uma melhora na qualidade da informação recuperada permitindo uma recuperação mais inteligente. Trabalhos futuros incluem a aplicação do método na coleção de documentos da Embrapa e a incorporação de uma ontologia do domínio de recursos naturais na base de conhecimento como, por exemplo, uma ontologia de recursos hídricos no Brasil.

6. REFERÊNCIAS

Apache Project. **Lucene**. Disponível em: <http://lucene.apache.org/java/docs/index.html>. Acesso em 12/05/ 2009.

BAEZA-YATES, R. A.; RIBEIRO NETO, B. A. **Modern Information Retrieval**. New York: ACM Press, 1st Edition, 1999. 513p.

LEITE, M. A. A.; RICARTE, I. L. M. Fuzzy information retrieval model based on multiple related ontologies. In: 20th IEEE INTERNATIONAL CONFERENCE ON TOOLS WITH ARTIFICIAL INTELLIGENCE. **Proceedings**… Ohio, USA, 2008.

LEITE, M. A. de A. **Modelo *Fuzzy* para Recuperação de Informação Utilizando Múltiplas Ontologias Relacionadas**, 2009. Tese de doutorado. Universidade Estadual de Campinas. Campinas. 183p. No prelo.

MANNING, C. D.; SCHUTZE, H. **Foundations of statistical natural language processing**. Cambridge: MIT Press, 1999. 680 p.

SISGA. **Mapa do Clima no Brasil**. Disponível em: <http://www2.inf.furb.br/sisga/educacao/ensino/mapaClima.php>. Acesso em 12/05/ 2009.