Expansão de busca utilizando vocabulário controlado

Marcos Aparecido Marinho Seixas[1]
Carlos Miguel Tobar Toledo[1]
Maria Fernanda Moura[2]

Introdução

No site da Agência de Informação Embrapa, ou simplesmente Agência, para facilitar a localização de informações desejadas pelos usuários, além de outros recursos de navegação, implantou-se um serviço de busca automática (CRUZ, 2003). Esse serviço de busca utiliza uma ferramenta open source chamada swish-e, que realiza a busca e a indexação dos hipertextos (no caso, páginas estáticas). A ferramenta de busca trata de assuntos gerais relacionados à agropecuária e com público variado, tais como publicações de cunho científico, publicitário, jornais, etc. A busca atual é realizada apenas nos metadados (SOUZA et al., 2004) dos recursos disponíveis no site, que são catalogados a partir do padrão Dublin Core. Dessa forma, os metadados obedecem a um padrão definido, e o assunto dos documentos, as palavras-chaves e as categorias agrícolas obedecem constantemente a um thesaurus – especificado junto ao metadado. Porém, toda essa qualificação prévia dos dados nem sempre é explorada pelo usuário final. Muitas vezes, as expressões de busca, definidas pelo usuário, poderiam ser expandidas, de acordo com os thesaurus utilizados para melhorar as respostas obtidas, seja na precisão ou na cobertura destas. Além disso, com uma possível expansão dessas buscas, seria interessante, além de mostrar, individualmente, o percentual de relevância das respostas, mostrá-lo também em uma visualização hierárquica.

[1] *Pontifícia Universidade Católica de Campinas (PUCCAMP), marcosams@cnptia.embrapa.br; tobar@puc-campinas.edu.br*
[2] *Embrapa Informática Agropecuária, fernanda@cnptia.embrapa.br*

Assim, o objetivo dessa proposta é expandir as expressões de busca especificadas pelo usuário, de modo a abrir o leque de resultados, melhorando especialmente a cobertura destes, conferindo uma maior precisão nos resultados das buscas.

Material e métodos

A evolução da ferramenta de busca, para incorporar a expansão, foi possibilitada pelo reúso da implementação de um Web Service, que permite acessar o banco de vocabulário controlado da Agência (SOUZA et al., 2010).

Conforme ilustrado na Figura 1 a), o usuário pode escolher se quer expandir a busca. A expansão automática procura no vocabulário controlado (via Web Service), pelos termos que devem ser usados no lugar do termo indicado ("use") ou dos termos que podem ser usados em seu lugar ("used for") e já os acrescenta à expressão de busca com o uso do operador "OR". Além disso, se o usuário desejar pode acrescentar termos relacionados ("related to – RT") aos termos de sua expressão de busca, podendo escolhê-los na relação apresentada. Supondo que o usuário escolheu o termo "Silo" (Figura 1), então a expressão é expandida para a forma apresentada na parte a) automaticamente, e a parte b) apresenta a expressão com os "related to" incluídos pelo usuário. A Figura 2 ilustra os resultados obtidos no caso de o usuário ter utilizado apenas a expressão "silo" e depois com a expansão escolhida na Figura 1 b). Nota-se, subjetivamente,

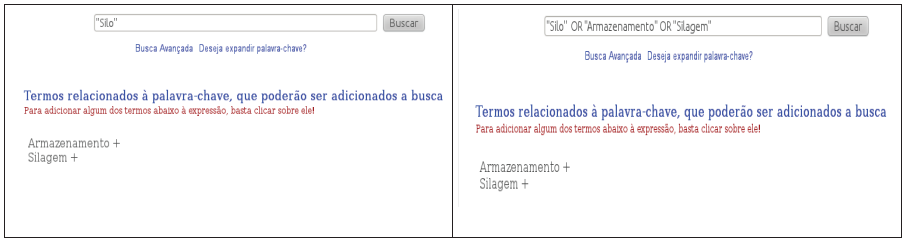

Figura 1. Expansão semântica de palavras-chave: a) usuário escolhe se quer expandir a busca; b) termos USE UF e RT.

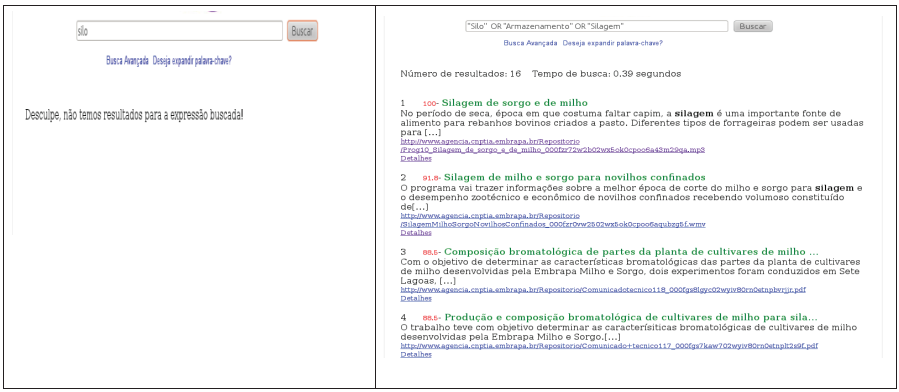

Figura 2. Comparativo dos resultados das buscas.

A evolução da ferramenta de busca, para incorporar a expansão, foi possibilitada pelo reúso da implementação de um Web Service, que permite acessar o banco de vocabulário controlado da Agência (SOUZA et al., 2010).

Conforme ilustrado na Figura 1 a), o usuário pode escolher se quer expandir a busca. A expansão automática procura no vocabulário controlado (via Web Service), pelos termos que devem ser usados no lugar do termo indicado (“use”) ou dos termos que podem ser usados em seu lugar (“used for”) e já os acrescenta à expressão de busca com o uso do operador “OR”. Além disso, se o usuário desejar pode acrescentar termos relacionados (“related to – RT”) aos termos de sua expressão de busca, podendo escolhê-los na relação apresentada. Supondo que o usuário escolheu o termo “Silo” (Figura 1), então a expressão é expandida para a forma apresentada na parte a) automaticamente, e a parte b) apresenta a expressão com os “related to” incluídos pelo usuário. A Figura 2 ilustra os resultados obtidos no caso de o usuário ter utilizado apenas a expressão “silo” e depois com a expansão escolhida na Figura 1 b). Nota-se, subjetivamente, uma boa melhora, resultado da busca, utilizando-se efetivamente a qualificação imposta aos metadados.

Resultados e discussão

Para validar objetivamente o processo, tem-se criado bases de dados artificiais, de modo a reproduzir o uso do vocabulário controlado tanto nos metadados quanto ao longo do textos. Por exemplo, em textos que tratam a cultura de cana-de-açúcar, criam-se novos textos que incorporam, em seus metadados e em seu corpo, os termos relacionados ao termo "cana--de-açúcar" e seus consequentes: cana-de-açúcar, canavieira, Saccharum officinarum, canavial, soca, soqueira, cultivo da soca. Dessa forma, tem-se uma base de textos que permite obter resultados de busca, para diferentes expressões de busca, com documentos relevantes e irrelevantes, considerando-se como documentos relevante só conjunto de documentos que contém, pelo menos, um dos termos relacionados ao assunto da expressão de busca, notado por DR.

Como trabalho futuro, será realizada uma análise da revocação e precisão das buscas em relação às expressões supostamente especificadas pelo usuário e às expressões com expansões aleatoriamente selecionadas. A revocação é a relação entre o número de documentos recuperados que efetivamente contém os termos da expressão de busca (tp – true positive) e o número total de documentos recuperados, e a precisão corresponde à relação entre tp e a cardinalidade de DR. A hipótese, estatisticamente esperada, é que se tenha em média uma melhor revocação para os resultados de expressões de busca expandidas.

Referências

CRUZ, S. A. B. da; CCRUZ, S. A. B. da. **Implantação de um serviço de busca em site da WWW. Campinas**: Embrapa Informática Agropecuária, 2003. 8 p. (Embrapa Informática Agropecuária. Comunicado técnico, 50).

SOUZA, M. I. F.; MOURA, M. F.; SANTOS, A. D. dos. **Estudo comparativo entre os metadados da Agência de Informação Embrapa e do acervo documental do AINFO**. Campinas: Embrapa Informática Agropecuária, 2004. 10 p. (Embrapa Informática Agropecuária. Comunicado técnico, 66).

SOUZA, M. I. F.; ALVES, M. das D. R.; QUEIROS, L. R.; SANTOS, A. D. dos; OLIVEIRA, L. H. M. de. Representação descritiva e temática no Sistema Agência de Informação Embrapa: controle de vocabulário. **Transinformação**, Campinas, v. 22, n. 1, p. 61-75, jan./abr. 2010.