Busca usando sinônimos no Ainfo-Consulta

Rogério Lecarião Leite[1]
Edmilson José Mangueira Carvalho[1]
Glauber José Vaz[2]

O AINFO é um sistema para automação de bibliotecas e recuperação de informações que permite a gestão da informação técnico-científica. É utilizado pelas bibliotecas da Embrapa, por organizações estaduais de pesquisa agropecuária e por outras instituições públicas e privadas (EMBRAPA INFORMÁTICA AGROPECUÁRIA, 2011a). Atualmente, o AINFO 6 é composto por três módulos: Ainfo-Gestor, responsável pelo registro e pelo controle de acervos, Ainfo-Digital, que armazena publicações digitais, e Ainfo-Consulta, que permite a realização de buscas de informações textuais baseadas nos registros do sistema e pode ser acessado sem necessidade de cadastro por qualquer terminal que possua conexão com a internet. A atual versão do Ainfo-Consulta não possui recurso de expansão de consulta por sinônimos, mas é interessante que, em uma busca com a utilização de determinada palavra, o resultado seja próximo àquele obtido em uma busca realizada com termo equivalente ou relacionado a esse termo. O foco deste trabalho é na expansão de consultas com termos obtidos por meio do Thesaurus Brasileiro de Agricultura - Thesagro (BRASIL, 1999) para melhorar os resultados do Ainfo-Consulta.

Para ilustrar essa situação, foram realizadas consultas na atual versão do Ainfo-Consulta (EMBRAPA INFORMÁTICA AGROPECUÁRIA, 2011b) com os termos ‘mandioca’, ‘manihot esculenta’, nome científico da mandioca, e ‘aipim’, outro nome usado para a palavra ‘mandioca’. Na Tabela 1, são exibidas as consultas realizadas por um usuário, associando os termos e operadores utilizados à quantidade de documentos recuperados.

[1] *Universidade Estadual de Campinas - Faculdade de Tecnologia, {edmilsonjmc, rogeriolI}@cnptia.embrapa.br*
[2] *Embrapa Informática Agropecuária, glauber@cnptia.embrapa.br*

Tabela 1. Busca por sinônimos no Ainfo-Consulta

Termo e operadores utilizados	Recuperados
mandioca	39324
“manihot esculenta”	4117
aipim	333
mandioca OU “manihot esculenta” OU aipim	39447

Observa-se claramente uma grande perda de informação que pode ser relevante. Com a exploração das relações entre termos apresentadas no Thesagro, o resultado de uma consulta utilizando individualmente qualquer um dos termos considerados traz 39.447 registros, equivalente ao resultado obtido pela consulta ‘mandioca OU “manihot esculenta” OU aimpim’. No entanto, na última versão do Ainfo-Consulta, a quantidade de registros recuperados por consultas individuais aos termos ‘mandioca’, ‘manihot esculenta’ e ‘aipim’ são 39.324, 4.117 e 333, respectivamente.

Portanto, se são considerados relevantes todos os documentos que apresentam sinônimos de palavras pesquisadas, apenas 10,44% dos documentos relevantes são recuperados na consulta a ‘manihot esculenta’ e, pior ainda, 0,84% na consulta a ‘aipim’.

O Thesagro especifica relações de equivalência USE (utilize-se) e UF (utilize-se para) entre termos justamente para indicar sinônimos. E essas relações são remissivas. Na Figura 1, são ilustradas as entradas no *thesaurus* para ‘MANDIOCA’, ‘AIPIM’ e ‘MANIHOT ESCULENTA’. Além das relações USE e UF, há outras que podem ser exploradas

	MANDIOCA
AIPIM	*UF AIPIM*
USE **MANDIOCA**	*UF MACAXEIRA*
	UF MANIHOT ESCULENTA
	UF MANIHOT UTILISSIMA
MANIHOT ESCULENTA	**BT** TUBERCULO
USE **MANDIOCA**	NT MANDIOCA BRAVA
	NT MANDIOCA MANSA
	RT TAPIOCA
	RT MANIVA

Figura 1. Exemplos de descritores no Thesagro.

futuramente, como as hierárquicas BT e NT, que indicam generalizações e especializações, e as de associação RT.

De acordo com o Thesagro, o termo AIPIM é direcionado para MANDIOCA, assim como MANIHOT ESCULENTA. O termo MANDIOCA, por sua vez, é utilizado para representar os termos AIPIM, MACAXEIRA, MANIHOT ESCULENTA e MANIHOT UTILISSIMA. Uma consulta a 'aipim', portanto, pode ser facilmente expandida para uma consulta que envolve os termos 'mandioca', 'macaxeira', 'manihot esculenta' e 'manihot utilissima'.

Para sua implementação, desde as primeiras versões, o Ainfo-Consulta é construído com o Apache Solr, uma plataforma de busca de código aberto (THE APACHE SOFTWARE FOUNDATION, 2011; SMILEY; PUGH, 2009). O Solr utiliza a biblioteca de recuperação de informação de alto desempenho e escalável Apache Lucene (GOSPODNETIC; HATCHER, 2005). Essas tecnologias já possuem recursos que permitem a utilização de sinônimos nas consultas e também ordenam os documentos recuperados de acordo com sua relevância.

Portanto, o Ainfo-Consulta, em sua versão atual, trata apenas os termos fornecidos explicitamente pelos usuários nas consultas, não oferecendo suporte para uma busca envolvendo os sinônimos desses termos, o que pode representar a perda de muitos documentos relevantes. O Solr e o Lucene fornecem recursos que possibilitam a exploração de sinônimos nas buscas e o Thesagro fornece um excelente conjunto de termos do domínio agrícola, identificando sinonímias que são úteis para a expansão das consultas no Ainfo-Consulta. Posteriormente, será possível trabalhar com pesos diferentes de relevância para cada termo acrescentado à consulta inicial, conforme suas relações, que podem envolver, além das equivalências, as hierárquicas e as de associação. Também será possível acrescentar descritores de outros thesaurus e novos recursos, como sugestões de termos parecidos e buscas de acordo com os perfis dos usuários.

Referências

THE APACHE SOFTWARE FOUNDATION. **Apache Solr**. Disponível em: <http://lucene.apache.org/solr>. Acesso em: 24 out. 2011.

BRASIL. Secretaria de Desenvolvimento Rural. **Thesagro**: thesaurus agrícola nacional. Brasília, DF: SDR, Cenagri, 1999. 242 p.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Ainfo**. 2011a. Disponível em: <http://www.ainfo.cnptia.embrapa.br>. Acesso em: 24 out. 2011.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Ainfo-Consulta**. 2011b. Disponível em: <http://ainfo.cnptia.embrapa.br/consulta>. Acesso em: 24 out. 2011.

GOSPODNETIC, O.; HATCHER, E. **Lucene in action**. Greenwich: Manning, 2005. 421 p.

SMILEY, D.; PUGH, E. **Solr 1.4 Enterprise Search Server**. Birmingham, UK: Packt Publishing, 2009. 317 p.