Boletim de Pesquisa 30
e Desenvolvimento

ISSN 1677-9266
Dezembro, 2011

Uma estratégia interativa para a expansão de expressões de busca utilizando vocabulário controlado

ISSN 1677-9266
Dezembro, 2011

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Informática Agropecuária
Ministério da Agricultura, Pecuária e Abastecimento

Boletim de Pesquisa e Desenvolvimento 30

Uma estratégia interativa para a expansão de expressões de busca utilizando vocabulário controlado

Maria Fernanda Moura
Marcos Aparecido Marinho Seixas
Carlos Miguel Tobar Toledo

Campinas, SP
2011

Embrapa Informática Agropecuária
Av. André Tosello, 209 - Barão Geraldo
Caixa Postal 6041 - 13083-886 - Campinas, SP
Fone: (19) 3211-5700 - Fax: (19) 3211-5754
www.cnptia.embrapa.br
sac@cnptia.embrapa.br

1ª edição on-line 2011

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Informática Agropecuária

Moura, Maria Fernanda.

Uma estratégia interativa para a expansão de expressões de busca utilizando vocabulário controlado / Maria Fernanda Moura, Marcos Aparecido Marinho Seixas, Carlos Miguel Tobar Toledo. - Campinas : Embrapa Informática Agropecuária, 2011.

24 p. : il. - (Boletim de pesquisa e desenvolvimento / Embrapa Informática Agropecuária, ISSN 1677-9266 ; 30).

1. Recuperação de informação. 2. Expansão de busca, thesaurus. 3. Vocabulário controlado. I. Seixas, Marcos Aparecido Marinho.II.Toledo, Carlos, Miguel Tobar.III. Embrapa Informática Agropecuária. IV. Título. V. Série.

CDD 025.3 (21. ed.)

Sumário

Uma estratégia interativa para a expansão de expressões de busca utilizando vocabulário controlado

Maria Fernanda Moura[1]
Marcos Aparecido Marinho Seixas[2]
Carlos Miguel Tobar Toledo[3]

Resumo

Neste trabalho apresenta-se a evolução do sistema de busca da Agência de Informação Embrapa, a partir de uma estratégia interativa de expansão de expressões de busca. Embora a Agência prime pela oferta de informação altamente qualificada, seus usuários não utilizam todo esse potencial, por desconhecer seus detalhes. A estratégia proposta permite que o usuário especifique sua expressão de busca e o sistema sugira formas de expandi-la, deixando a escolha final para o usuário. Os experimentos realizados foram baseados em possibilidades automaticamente geradas e medidos de forma objetiva, considerando-se a precisão e a cobertura das buscas realizadas. Como esperado, os experimentos mostram que o uso de uma melhor qualificação das expressões permite obter resultados mais precisos e que a evolução da ferramenta atingiu seu objetivo.

Termos para indexação: Recuperação de informação, expansão de busca, thesaurus, vocabulário controlado

[1] *Doutora em Ciências de computação e matemática computacional, Pesquisadora da Embrapa Informática Agropecuária, Caixa Postal 6041, Barão Geraldo, 13083-886, Campinas, SP, fernanda@cnptia.embrapa.br*
[2] *Estagiário da Faculdade de Engenharia de Computação da PUC de Campinas, Jd. das Universidades, Caixa-Postal 317, Barão Geraldo, 13020-904, Campinas, SP, touppermg@gmail.com*
[3] *Doutor em Engenharia Elétrica e de Computação, Professor titular da PUC de Campinas, Faculdade de Engenharia de Computação, Campinas, SP, tobar@puc-campinas.edu.br*

An interactive strategy for the expansion of search expressions using controlled vocabulary

Abstract

This paper presents the evolution of the Embrapa Information Agency search system, through an interactive strategy of query expansion. Although this search system aims to offer high qualified information, its users do not use this entirely potencial, because they do not know such details. The proposed strategy allows the user to specify the desired query and the system suggests query expansion alternatives; the decision is up to the user. The experiments were carried out based on automatically generated query expansion and were evaluated through the objective measures precision and recall. As expected, the experiments showed that the use of a better qualification from the expressions allows to obtain more precise results, consequentely the tool evolution reached its goal.

***Index terms**: Information retrieval, query expansion, thesaurus, controlled vocabulary*

Introdução

O advento da internet e a possibilidade da digitalização dos documentos causaram um aumento massivo da quantidade de informação disponível, surgindo o desafio de se encontrar a informação desejada, o que, frequentemente, é uma tarefa difícil e exaustiva. Na web, as páginas são acessadas por meio de um endereço. Para que o usuário tenha acesso a informação que ele necessita, é preciso que ele tenha o endereço da página ou utilize um sistema de busca, e tenha, como resposta, vários links para páginas que contenham informação relevante relacionada ao assunto.

Os mecanismos de busca possuem uma interface, normalmente uma página web, que é utilizada pelos usuários para efetuar a pesquisa em uma base de dados ou na internet. O mecanismo deve fornecer meios para que o usuário formule a sua consulta com palavras-chave, a qual é recebida e transmitida para a ferramenta ou motor de busca propriamente dito. Esse é um programa que localiza, entre os vários itens indexados na base de dados, aqueles que devem constituir a resposta. A ferramenta também é responsável pela ordenação dos resultados, de maneira que os mais relevantes apareçam em primeiro lugar. Os resultados mostrados contêm uma lista de descrições de sites e seus links.

Um dos problemas mais comuns enfrentados pelos usuários, ao utilizarem ferramentas de recuperação de informação, é o desconhecimento da forma em que os documentos estão catalogados e indexados nos repositórios de busca que os armazenam. No entanto, mesmo quando o usuário da ferramenta possui conhecimento sobre um determinado domínio, as informações armazenadas podem não corresponder à forma como as informações estão representadas.

Em geral, os usuários têm dificuldades para construir expressões de buscas objetivas, assim, os resultados obtidos em grande parte das buscas são irrelevantes, e a consulta é refeita várias vezes, o que pode ser uma tarefa frustrante e exaustiva. Nesse contexto, a expansão de consultas se propõe a manipular os termos da consulta inicial para torná-la mais eficaz, seja diminuindo a quantidade de informações irrelevantes, seja influenciando no processo de consulta para que resultados relevantes não sejam desprezados (TOMONARI et al., 2005).

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*, Swish-e (SWISH-E, 2011), que realiza as buscas e indexações dos hipertextos (no caso, páginas estáticas). Essa ferramenta indexa os hipertextos existentes em um repositório padrão e, posteriormente, de acordo com a demanda dos usuários, realiza as buscas pelos documentos que contém os termos repassados pela aplicação principal, que é responsável pela interação com o meio externo. 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*. Os metadados obedecem a um padrão definido e o assunto dos documentos, as palavras-chave e as categorias agrícolas obedecem, constantemente, a um *thesaurus* – especificado junto ao metadado.

Muitas vezes as consultas realizadas por essa ferramenta poderiam retornar resultados mais apurados quanto à relevância, mas como a maioria de seus usuários desconhece o padrão de metadados utilizado, isso não ocorre. Logo, se estivesse presente na ferramenta um sistema que pudesse auxiliar o usuário na formulação das suas expressões de busca, consequentemente haveria uma exploração mais qualitativa dos dados dos hipertextos. Especialmente porque a filosofia de construção das Agências é baseada na qualificação da informação consumida pelos usuários e, para atingir esse objetivo, os metadados são qualificados por thesaurus e termos controlados, que ainda passam por auditoria antes de sua liberação (ROSA et al., 2009).

Assim, o objetivo dessa proposta é explorar a qualificação dos dados dos hipertextos da Agência de Informação via expansão das expressões de busca definidas pelos usuários. Para que essa proposta seja viabilizada, será utilizado um vocabulário controlado (*thesaurus*) brasileiro, Thesagro (BINAGRI, 2011), para realizar a expansão da expressão de busca especificada pelo usuário. Com isso, tem-se o intuito de obter uma maior precisão e uma cobertura mais abrangente dos resultados para as consultas dos usuários.

Material e métodos

Como este trabalho apresenta uma evolução de uma ferramenta de software já existente, nos próximos itens explicam-se os materiais e métodos desse tipo de problema: o motor de busca utilizado e sua adaptação ao processo a ser evoluído; a integração do uso de um vocabulário controlado à ferramenta, como interpretar as relações desse vocabulário e o *web service* que permite consultá-lo; as estratégias de expansão de consulta adotadas; as estratégias de avaliação das expansões; e, as métricas para avaliar o desempenho essas expansões.

O sistema foi implementado utilizando a linguagem de programação PHP 5 (PHP, 2011).

2.1 Ferramenta de busca e indexação Swish-E

Foi desenvolvida interface para que a Swish-e interagisse com programas feitos em C, Perl, Python , PHP, entre outras linguagens de programação. Ela utiliza um arquivo de configurações (CONFIG na Figura 1), onde são definidos os parâmetros para a indexação e busca dos documentos, tais como: endereço do repositório, metadados de interesse e tipos de arquivos.

No sistema Agência, as tarefas de indexação (Figura 1), recuperação (Figura 2) e ordenação das respostas são realizadas pela swich-e, baseando-se apenas no conteúdo, sem fazer uso de quaisquer ferramentas auxiliares. Os documentos são indexadods por meio de arquivos invertidos que contêm para cada palavra P: (1) sua frequência no arquivo invertido; (2) a lista de URLs de todos os documentos Di nos quais P ocorre; (3) a lista de frequências de P em cada Di ; (4) a lista de posições de P em cada Di e (5) a lista de marcadores HyperText Markup Language (HTML) associados a P em cada Di. A ferramenta recupera documentos por meio de buscas booleanas realizadas no arquivo invertido. A ordenação das respostas obtidas é realizada heuristicamente considerando a medida term frequency inverse document frequency (tf-idf) (BAEZA-YATES; RIBEIRO-NETO, 1999), a posição do termo na consulta e no documento, e a proximidade dos termos da consulta dentro do documento.

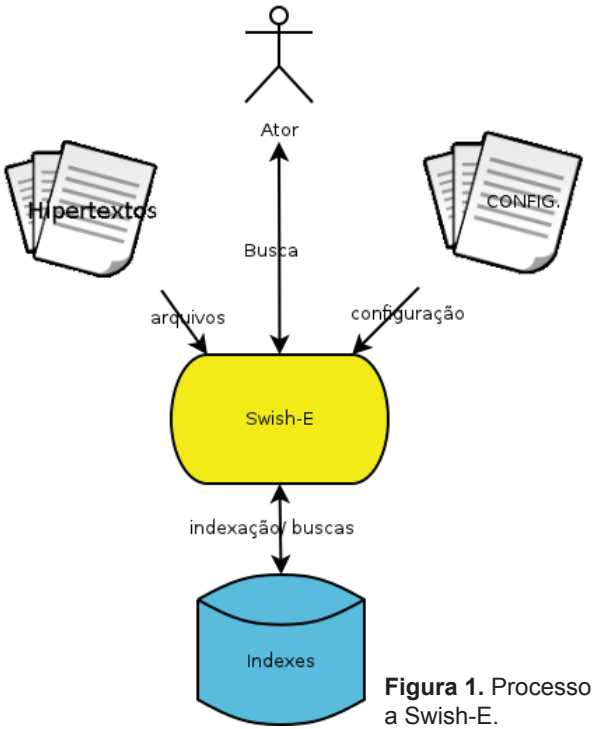

Figura 1. Processo de indexação e busca com a Swish-E.

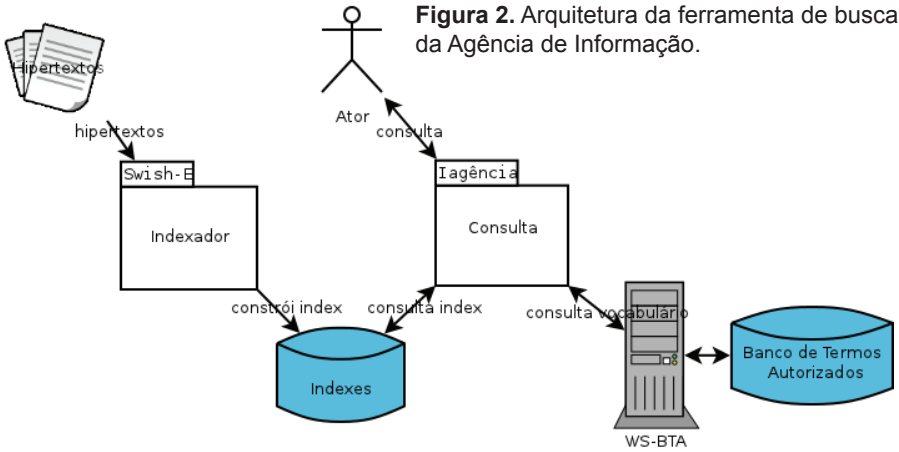

Figura 2. Arquitetura da ferramenta de busca da Agência de Informação.

As buscas podem ser realizadas utilizando expressões das mais variadas formas, por exemplo:

- Expressão simples: Cana de açúcar
- Operadores booleanos: 'Cana AND Plantio'
  - 'Etanol OR Álcool'
  - 'Produção AND (NOT Etanol)'
- Curingas: * - quando substituído na palavra-chave, ele pode representar zero ou mais caracteres;
  - Ex: 'Can*'
  - ? - quando substituído na palavra-chave, ele só pode representar exatamente um único caractere.
  - Ex: 'Can?'
- Busca com metadados: 'title: Época de plantio'
  - 'title: Etanol AND description: preço '

2.2 Vocabulário controlado

O vocabulário controlado utilizado para expansão da consulta foi basicamente o THESAGRO, que é o *thesaurus* mantido pela Biblioteca Nacional de Agricultura (BINAGRI), órgão da Secretaria de Executiva do Ministério da Agricultura, Pecuária e Abastecimento. O Thesagro é disponível gratuitamente na Internet, com uma estrutura enriquecida, que contém 9.351 termos, que podem ser referenciados por meio de três relações semânticas, ilustradas por meio dos termos exemplos "cana-de-açúcar" na Figura 3 (a e b).

(a)

CANA DE AÇÚCAR

UF CANAVIEIRA
BT PLANTA PRODUTORA DE ACUCAR
NT CANA FORRAGEIRA
NT CANA PLANTA
RT SACCHARUM OFFICINARUM
RT CANAVIAL
RT SOCA

(b)

CANAVIEIRA

USE CANA DE AÇÚCAR

Figura 3. Estrutura da hierarquia das relações do vocabulário controlado THESAGRO.

- **Relação de associação** - *RELATED TERM (RT)*: o termo relacionado é a relação que estabelece a associação entre um Descritor cujo significado se relaciona semanticamente com outro Descritor, mas sem nenhuma ligação hierárquica entre si.
- **Relação de equivalência USE** (Utilize-se): é empregada para indicar qual o Descritor autorizado para ser usado em lugar do atual. A relação não aparece no exemplo, ela apareceria se o termo fosse “canavieira”, pois é a relação inversa à UF.
- **Relação de equivalência UF** - USE FOR (Utilize-se para): é a relação empregada para indicação do Descritor não autorizado, em favor do Descritor autorizado, sob o qual a informação é colocada.
- **Relações hierárquicas** -
  - *BROADER TERM (BT)*: ou termo genérico, que engloba o conceito deste, é a relação empregada para indicar o Descritor mais amplo, mais abrangente.
  - *NARROWER TERM (NT)*: ou termo específico, é empregada para indicar Descritores mais específicos.

Para este trabalho foram utilizadas apenas as relações de USE, UF e RT. Isso pelo o banco de termos autorizado da Agência (ROSA et al., 2009) ter sido especificado para utilizar essas relações como guias de catalogação, e, desta forma apenas essas relações terem sido requisitadas para a evolução da ferramenta de busca. A ferramenta que provê o serviço de recuperação dessas relações é descrita no próximo item

2.2.1 *Web Service* – Banco de Termos Autorizados

A consulta ao vocabulário controlado é feita via utilização de um serviço disponibilizado por um *web service* desenvolvido na Embrapa Informática Agropecuária (QUEIROS et al., 2009), denominado WS-BTA. Esse *web service* foi implementado utilizando a tecnologia REST v. 1.0.1 (REST, 2011), que apresenta as seguintes características:

- Um protocolo cliente/servidor sem estado: cada mensagem HTTP contém toda a informação necessária para compreender o pedido. Como resultado, nem o cliente e nem o servidor necessitam gravar nenhum estado das comunicações entre mensagens. Na prática, muitas aplicações

baseadas em HTTP utilizam *cookies* e outros mecanismos para manter o estado da sessão (algumas destas práticas, como a reescrita de URLs, não são permitidas pela regra do REST).
- Um conjunto de operações bem definidas que se aplicam a todos os recursos de informação: HTTP em si define um pequeno conjunto de operações, as mais importantes são POST, GET, PUT e DELETE. Com frequência, essas operações são combinadas com operações CRUD para a persistência de dados, onde POST não se encaixa exatamente nesse esquema.
- Uma sintaxe universal para identificar os recursos. No sistema REST, cada recurso é unicamente direcionado pela sua URI.
- O uso de hipermídia, tanto para a informação da aplicação como para as transições de estado da aplicação: a representação desse estado em um sistema REST são tipicamente HTML ou XML. Como resultado disso, é possível navegar com um recurso REST a muitos outros, simplesmente seguindo ligações sem requerer o uso de registros ou outra infraestrutura adicional.

O WS-BTA possui em seu banco de dados o mapeamento dos *thesaurus* NAL e THESAGRO, bem como os termos livres da Agência; e, o serviço oferecido é a disponibilização da consulta a esses *thesaurus*.

As consultas são realizadas em duas etapas:

1- Consulta ID, do termo a ser pesquisado pela palavra-chave;
2- A seguir, por meio do ID encontrado, é realizada a pesquisa ao termo. O sistema devolve um arquivo no padrão XML contendo uma hierarquia do termo consultado e suas relações.

2.2. Estratégias de expansão de consulta

Neste trabalho o principal objetivo era promover maior interação entre o usuário e a ferramenta de busca. O projeto não optou por uma estratégia automática de expansão da busca e sim por torná-la interativa. O usuário escolhe na interface se quer expandir a busca, neste caso, primeiramente os termos das relações USE e UF são incorporados aos termos da expres-

são de busca especificada pelo usuário e presentes no banco de termos autorizados. A nova expressão de busca é apresentada ao usuário, que é composta pela expressão original e uma disjunção booleana (OR) com os termos das relações UF e USE. Então, o usuário pode decidir se quer utilizar mais termos na sua busca, pois lhe são mostrados os termos RT em uma lista, onde ele pode selecionar os de seu interesse, conforme ilustração nas Figuras 4, 5 e 6.

Figura 4. Expressão de busca sem expandir.

Figura 5. Expressão de busca agregada com os termos das relações UF e USE.

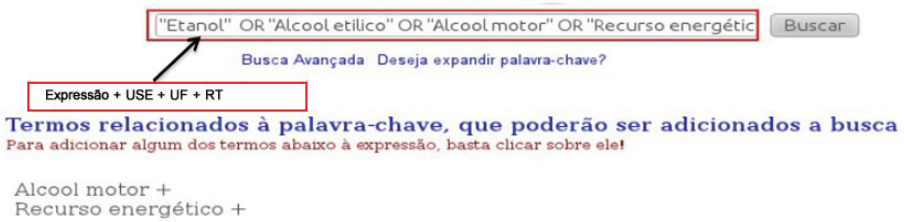

Figura 6. Expressão de busca agregada com os termos das relações UF, USE e RT.

Assim, as seguintes estratégias de busca, para validar as expansões, foram utilizadas:

1- Expressão original, sem expansão;
2- Expansão automática utilizando os termos relacionados USE e UF;
3- Expansão manual utilizando os termos relacionados USE, UF e RT.

Note que, independentemente da estratégia, o processo de expansão é caracterizado pela substituição de cada palavra-chave pertencente à consulta original por uma disjunção booleana (OR) da própria palavra-chave e seus termos relacionados fornecidos pelo WS-BTA.

2.3 Estratégia de avaliação

Para os experimentos, foi utilizada uma coleção de hipertextos de assuntos relacionados à cultura da cana-de-açúcar, obtidos da Agência de Informação da Cana-de-Açúcar, (http://www.agencia.cnptia.embrapa.br/gestor/cana-de-acucar/Abertura.html). Os hipertextos foram alterados, isto é, criados outros hipertextos a partir dos originais, para possibilitarem a avaliação da ferramenta. As alterações foram a inserção de palavras-chave pré-definidas pelo avaliador do sistema em vários campos de metadados diferentes (Figura 4), da seguinte forma:

- 10 hipertextos somente com os termos originais da busca (palavra-chave);
- 5 hipertextos somente com os termos USE e UF relacionados à palavra-chave;
- 5 hipertextos com os termos RT relacionados à palavra-chave;

Esses vinte textos foram considerados como os textos relevantes da coleção de hipertextos e os demais, os hipertextos irrelevantes; sendo que a base contém um total de 110 hipertextos.

Obtida a coleção de hipertextos a ser usada nos experimentos, elaboraram-se cinco consultas com os termos “etanol”, “plantio”, “doença”, “álcool” e “adubo” e seus relacionamentos:

- **TERMO**:”etanol”
  UF:”alcool etilico”
  RT:{“alcool motor”,”recurso energetico”}
- **TERMO**:”plantio”
  UF:”plantação”
  RT:{“densidade de plantio”,”espacamento”,”fileira dupla”,”plantadeira”,”pre plantio”,”epoca de plantio”}
- **TERMO**:”Doença”,
  UF:{”mal”,”patologia”}
  RT: {“diagnóstico”,”distúrbio”,”epidemiologia”,”erradicação”,”fisiopatolo-gia”, “hospedeiro”,”medicina”,”patogenicidade”,”profilaxia”,”sintoma”}
- **TERMO**:”alcool”
  RT:{“combustível”,”destilaria”}
- **TERMO**:”adubo”
  RT:{”adubacao”,”adubadeira”,”amonia”,”fertilizante”}

2.4 Métricas de avaliação

Métricas para medir a eficiência de um sistema de recuperação de informação são medidas usadas para calcular a habilidade desses sistemas em recuperar os documentos relevantes existentes em sua base de dados, em caso de necessidade de informação do usuário; e, as métricas mais importantes na avaliação de um mecanismo de recuperação de informação são a cobertura e a precisão (BAEZA-YATES; RIBEIRO-NETO, 1999).

Neste trabalho, mede-se a eficiência para cada estratégia de expansão adotada, usando as métricas precisão, cobertura e a medida-F (Figura 7b), para classificar os resultados das buscas foi utilizada uma estrutura conhecida como matriz de confusão, que é uma tabela contingência das frequencias de documentos recuperados, correta e incorretamente. A matriz de confusão representa quatro categorias: True Positives (TP), que são os hipertextos corretamente classificados como relevantes; e False Positives (FP), que referem-se aos hipertextos incorretamente classificados como relevantes; True Negatives, (TN) que correspondem aos hipertextos corretamente classificados como irrelevantes; e False Negative (FN) que

referem-se aos hipertextos incorretamente classificados como irrelevantes (Figura 7a).

	Real positivo	Real negativo
Predição positiva	TP	FP
Predição negativa	FN	TN

(a) Matriz de resultados

1. Precisão (P) = TP / TP + FP
2. Cobertura (C) = TP / TP + FN
3. Medida-F = 2 * (P * C) / (P + C)

(b) Definição das métricas

Figura 7. Métricas para avaliação da ferramenta de busca.

Na recuperação da informação, um valor ideal para a precisão é “um”, isso significa que todos os resultados recuperados pela busca são relevantes; porém essa medida não indica se todos os documentos relevantes foram recuperados. O valor ideal para a cobertura também é “um”, o que significa que todos os documentos relevantes foram recuperados pela busca; mas, esta nada diz sobre o número de documentos irrelevantes também recuperados. A cobertura pode ser facilmente melhorada com o decréscimo da precisão, e vice-versa, para avaliar os ganhos da eficiência.

Para os experimentos desse trabalho, com o uso da expansão da consulta, foi também utilizada a medida-F, que simboliza o comportamento harmônico das duas primeiras. A medida-F cresce não apenas quando a precisão cresce com um valor fixo para a cobertura ou vice-versa, mas ela cresce, também, quando os valores da precisão e cobertura se aproximam. Essa é uma característica importante e desejada para uma métrica de avaliação de sistema de recuperação de informação, pois ela avalia tanto o comportamento da precisão, como o da cobertura, e mais ainda, o comprometimento de uma dada estratégia para essas duas medidas.

Resultados e discussão

Nas Tabelas 1, 2 e 3 estão, respectivamente, os resultados dos experimentos realizados para as três estratégias de busca:

Tabela 1. Resultados das buscas usando a estratégia (1) - não expandir as expressões.

Termo	N° total de resultados	N° de resultados relevantes	Precisão	Cobertura	medida-F
Etanol	12	10	0,45	0,83	0,58
Doença	14	10	0,5	0,77	0,61
Plantio	13	10	0,5	0,77	0,61
Álcool	20	10	0,67	0,5	0,57
Adubo	12	10	0,67	0,83	0,74

Tabela 2. Tabela de resultados das buscas usando a estratégia (2) de expansão automática, de forma que somente os termos das relações USE e UF são agregados à expressão original.

Termo	N° total de resultados	N° de resultados relevantes	Precisão	Cobertura	medida-F
Etanol	17	15	0,75	0,88	0,81
Doença	19	15	0,75	0,79	0,77
Plantio	19	15	0,75	0,79	0,77

Tabela 3. Tabela de resultados das buscas, nessa estratégia de expansão são agregados à expressão original os termos dos relacionamentos USE, UF e RT.

Termo	N° total de resultados	N° de resultados relevantes	Precisão	Cobertura	medida-F
Etanol	22	20	1	0,91	0,95
Doença	24	20	1	0,83	0,91
Plantio	25	20	1	0,80	0,89
Álcool	26	15	1	0,58	0,73
Adubo	28	15	1	0,54	0,70

1 - Sem expansão
2 - Expansão automática utilizando os termos das relações USE e UF
3 - Expansão manual utilizando os termos das relações USE, UF e RT

Nas Figuras 8, 9, 10, 11, 12 e 13 estão, respectivamente, representados os gráficos dos resultados dos cálculos das métricas. Cada gráfico representa os experimentos realizados para cada uma das estratégias de expansão semântica das expressões de busca e os seus respectivos valores de precisão, cobertura e medida-F.

De acordo com os resultados apresentados nas Tabelas 1, 2 e 3, observa-se claramente que, quando apenas o termo principal é indicado para a expressão de busca, a precisão tem uma eficiência média (em torno de 50%) enquanto a cobertura é até bem razoável, mas o balanço harmônico entre as duas (medida F) mantém um comportamento médio. Quanto mais termos corretos são agregados à busca, mais a média harmônica se aproxima do valor 1, que é seu valor ideal.

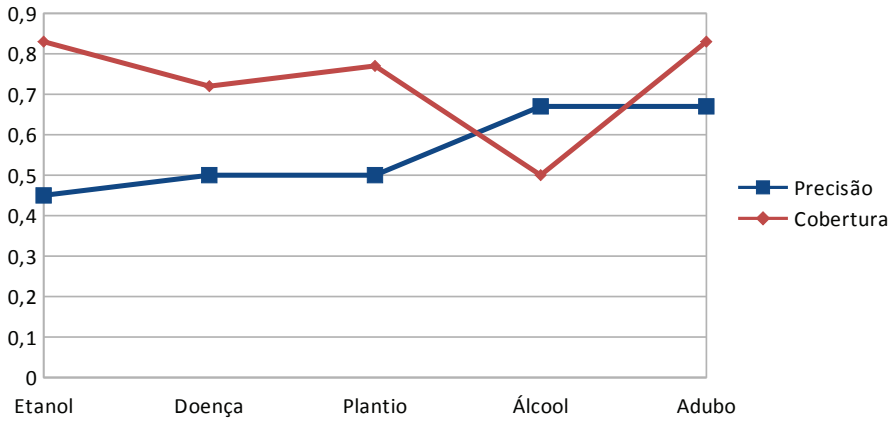

Figura 8. Gráfico dos resultados das buscas usando a estratégia de não expandir as expressões.

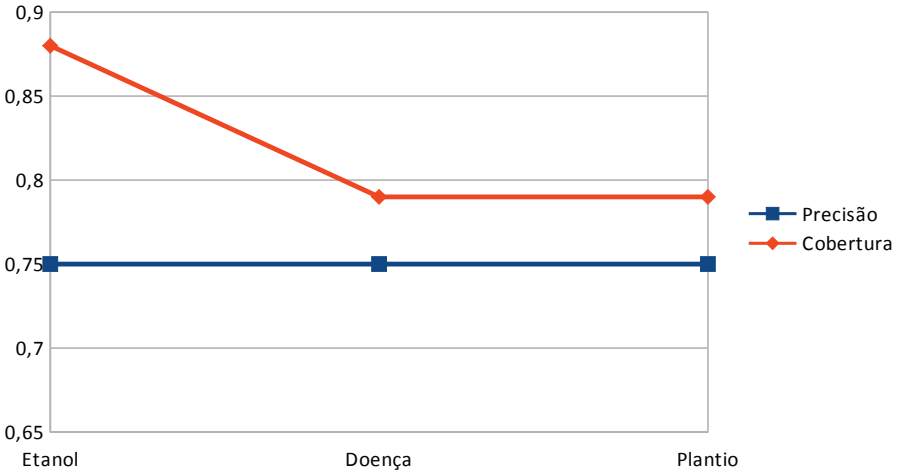

Figura 9. Gráfico da média harmônica (medida-F) entre a precisão e a cobertura, usando a estratégia de não expandir as expressões.

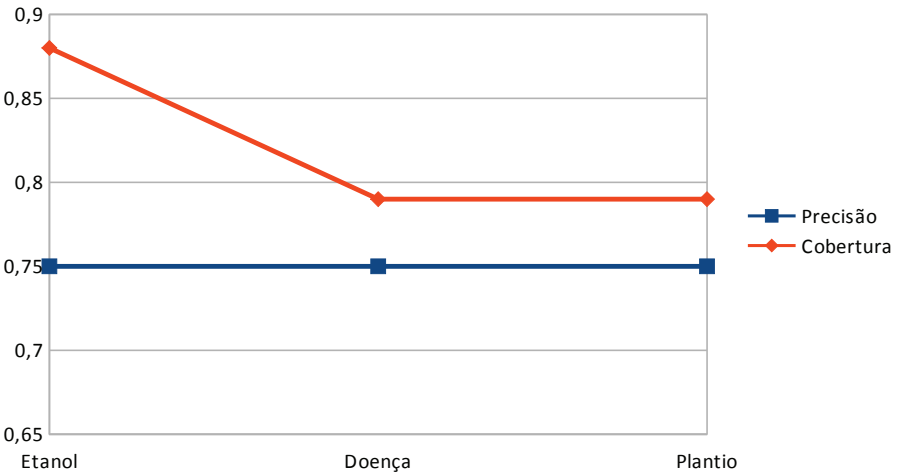

Figura 10. Gráfico dos resultados das buscas, nessa estratégia é feita a expansão automática, de forma que somente os termos das relações USE e UF são agregados à expressão original. Note que para essa estratégia existe somente três experimentos, porque as palavras-chave 'Álcool' e 'Adubo' não possuem esses relacionamentos.

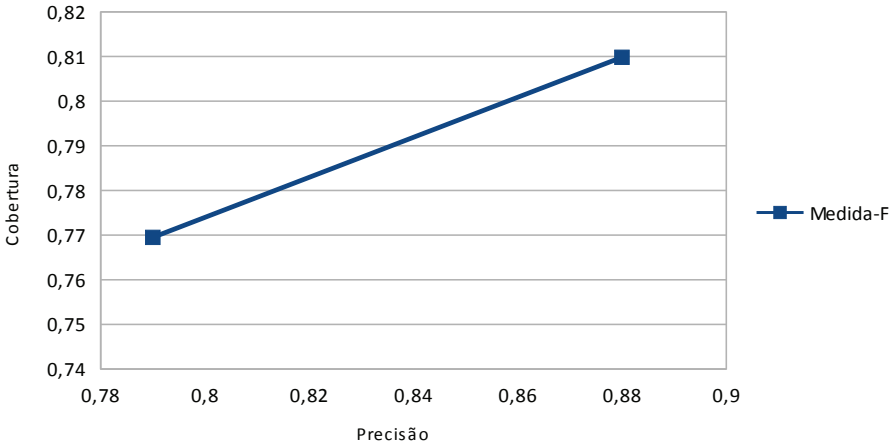

Figura 11. Gráfico da média harmônica (medida-F) entre a precisão e a cobertura. Nessa estratégia a expansão é feita automaticamente, de forma que somente os termos das relações USE e UF são agregados à expressão original.

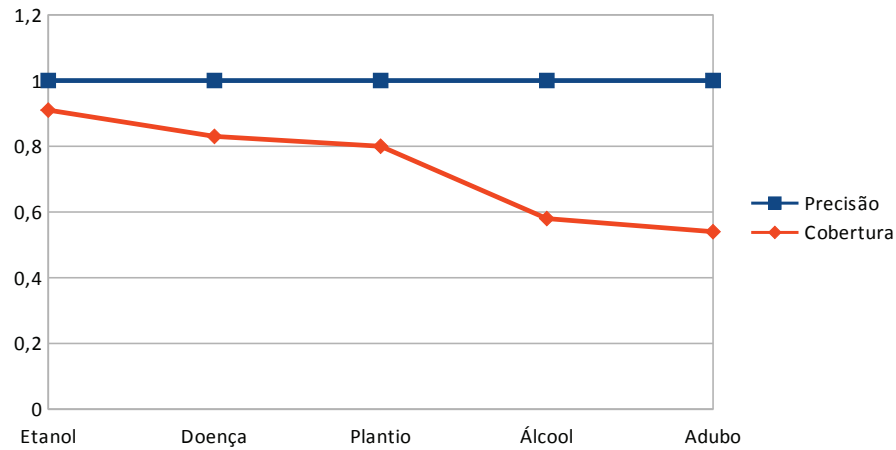

Figura 12. Gráfico dos resultados de buscas. Nessa estratégia de expansão, os termos dos relacionamentos USE, UF e RT são agregados à expressão original, os termos dos relacionamentos USE, UF e RT. Foi o experimento que apresentou o resultado mais relevante quanto à precisão, isso devido à quantidade de termos presentes na expressão de busca, o que ocasionou uma melhor exploração dos dados presentes nos hipertexto.

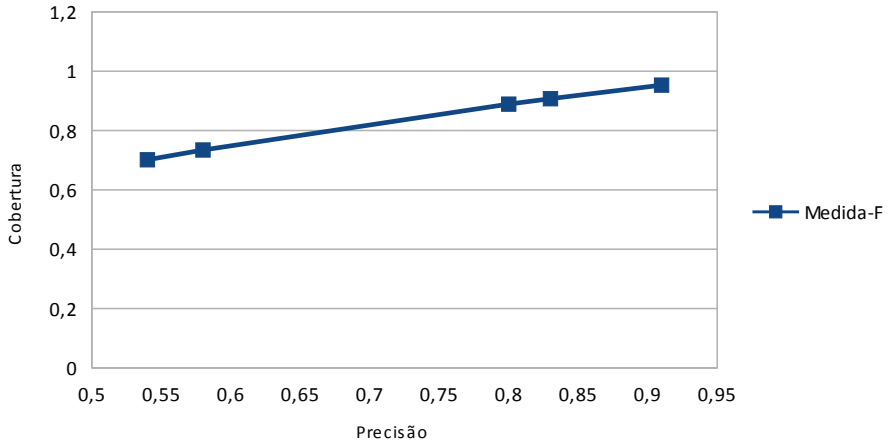

Figura 13. Gráfico da média harmônica (medida-F) entre a precisão e a cobertura. Nessa estratégia de expansão são agregados, à expressão original, os termos dos relacionamentos USE, UF e RT.

Como era de se esperar, as expressões que apresentam maior número de termos interessantes e corretamente utilizados no domínio de conhecimento, consequentemente presentes nos hipertextos mais relevantes da coleção, apresentam resultados mais precisos (Figura 6). Isso ocorre porque os documentos da coleção de hipertextos utilizada no experimento são forçadamente bem qualificados, apresentando os termos corretamente utilizados no domínio em seus metadados e no corpo dos hipertextos. Essa situação de alta qualificação é a esperada na Agência, pois ela ocorre obrigatoriamente no processo de catalogação dos documentos. Deste modo, mesmo que os termos não apareçam no corpo do documento, aparecerão, pelo menos, nos metadados.

Conclusões

Foi apresentado, neste artigo, um experimento que objetivamente mediu o impacto em recuperação de informação do processo de expandir semanticamente expressões de busca, realizada de forma automática ou manual.

A tarefa de expansão foi auxiliada por um vocabulário controlado que forneceu as palavras semanticamente relacionadas com os termos originais da consulta. Os experimentos comprovaram que, de uma forma geral, tal expansão sempre melhora a eficiência média sobre o conjunto total de documentos recuperados.

Ainda, embora este trabalho não apresente resultados de experimentos com medidas subjetivas, isto é, avaliados por usuários da ferramenta, a ferramenta atingiu seus objetivos de acordo com os requisitos de sua especificação, que correspondem a permitir a expansão da busca de forma interativa e com apoio de vocabulário controlado, de modo a recuperar com maior precisão os hipertextos da Agência, cuja qualificação tem base no mesmo vocabulário controlado utilizado na busca.

Para trabalhos futuros pretende-se incorporar a API de busca e indexação Lucene (APACHE, 2011) na ferramenta, não só para comparar a eficiência em relação à Swish-e, mas também para permitir expandir a busca para bases de dados – indexadas pela Lucene. Pode-se decidir por manter a Swich-e para os hipertextos e a Lucene para base de dados, dependendo dos resultados da comparação.

Agradecimentos

Os autores gostariam de registrar um agradecimento aos desenvolvedores do WS-BTA, Leonardo Queirós e Leandro Mendonça, pela disposição do código fonte e pela sua prestatividade na dissolução de dúvidas.

Referências

BAEZA-YATES; RIBEIRO-NETO. Modern Information Retrieval. Addison-Wesley, 1999.

BINAGRI (Brasil). **Biblioteca Nacional de Agricultura**. Disponível em: <http://snida.agricultura.gov.br:81/binagri/html/Cen_Thes1.html>. Acesso em: 10 ago. 2011.

CRUZ, 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). Disponível em: <http://ainfo.cnptia.embrapa.br/digital/bitstream/CNPTIA/10051/1/comtec50.pdf>. Acesso Acesso em: 30 maio 2008.

PHP - PHP manual. 2011. Disponível em : <www.php.net/manual>. Acesso em: 15 mar. 2011.

QUEIROS, L. R.; OLIVEIRA, L. H. M. de; SOUZA, M. I. F. **WS-BTA**. Versão 1.1.Campinas: Embrapa Informática Agropecuária, 2009.1 CD-ROM.

RESTFUL - RESTful Web Services. Disponível em: <http://www.oracle.com/technetwork/articles/javase/index-137171.html>. 15 dez. 2011.

ROSA, M. das D. A.; SOUZA, M. I. F.; QUEIROS, L. R. **Banco de termos** - ferramenta para controle e padronização de termos na Agência de Informação Embrapa. Campinas: Embrapa Informática Agropecuária, 2009. 7 p. (Embrapa Informática Agropecuária. Comunicado técnico,97). Disponível em: <http://ainfo.cnptia.embrapa.br/digital/bitstream/item/17790/1/cnptiact97_0.pdf>. Acesso em 17 dez. 2011.

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).

SWISH-e - simple web indexing system for humans - enhanced. Disponível em: <http://swish-e.org/>. Acesso em: 5 abr. 2011.

TOMONARI, M.; TERUHITO, K.; ATSUHIRO, T. ; JUN, A. Improving Web search by query expansion with a small number of terms. In: WORKSHOP MEETING, 5., 2005, Tokyo. **Proceedings**... Tokio: NTCIR, 2005. 8 p.

Informática Agropecuária

Ministério da
Agricultura, Pecuária
e Abastecimento

GOVERNO FEDERAL
BRASIL
PAÍS RICO É PAÍS SEM POBREZA

CGPE 9795