Uma metodologia para criação de *stop lists* em sistemas de recuperação de informação em domínios específicos

Henrique Tadeu Merjam Cardone[1]
Igor Jones Proença Marinho[1]
Glauber José Vaz[2]

Em um sistema de recuperação de informação (SRI), que é capaz de catalogar e recuperar documentos relevantes à consulta do usuário, nem todas as palavras presentes nos documentos são bons descritores para recuperá-los. Essas palavras irrelevantes que não possuem valor semântico e ocorrem com frequência significativa, como por exemplo, artigos, preposições e conjunções, são denominadas *stop words*. Sua remoção dos índices gerados em SRIs normalmente visam a: (i) diminuir o tamanho do índice; (ii) tornar mais rápidas as consultas a frases que envolvam *stop words*; e (iii) melhorar o ranking dos resultados (THE APACHE SOFTWARE FOUNDATION, 2012). Assim, é comum que durante a indexação dos documentos ou a realização da busca, um SRI recorra a uma lista de *stop words*, denominada *stop list*, na análise dos textos. Para isso, normalmente, considera-se uma lista pré-determinada de palavras que já são consideradas *stop words* nos diferentes idiomas presentes nos documentos. No entanto, a partir da análise das palavras mais comuns no conjunto de documentos catalogados, é possível criar uma *stop list* mais adequada ao contexto. O objetivo deste trabalho é propor uma metodologia para a construção de *stop lists* baseada no domínio da aplicação. Aqui, consideramos o uso da metodologia na construção da *stop list* para o Ainfo-Consulta, SRI que possibilita a realização de pesquisas nos acervos impressos e digitais de toda a Embrapa (EMBRAPA INFORMÁTICA AGROPECUÁRIA, 2012).

[1] Faculdade de Tecnologia da Unicamp, {htcardone, igorj27}@gmail.com
[2] Embrapa Informática Agropecuária, glauber.vaz@embrapa.br

O Ainfo-Consulta utiliza a plataforma de código aberto Apache Solr, que oferece, de maneira simples, os recursos do motor de busca Apache Lucene. Uma outra ferramenta de código aberto que auxilia na construção de SRIs é a Luke (*Lucene Index Toolbox*), que possibilita a análise de índices construídos com a ferramenta Lucene. Ela oferece, dentre outras funcionalidades, um ranking dos termos encontrados nos documentos indexados segundo sua frequência. A Figura 1 exibe uma tela do Luke que representa a situação de um índice gerado a partir dos dados do Ainfo-Consuta. Ela mostra, à direita, os termos da posição 34 à 44 referentes ao campo 'resumo'. O termo 'objetivo', por exemplo, é o 35º mais frequente neste campo e está presente em 40.844 documentos. Outros oito campos podem ser observados à esquerda da Figura 1, que exibe a quantidade de termos diferentes contidos em cada campo e a porcentagem que este ocupa no índice. O campo 'resumo', por exemplo, conta com 63.760.399 termos diferentes e ocupa 33,95% do índice. O parâmetro no centro da tela indica que devem ser exibidos os 1.000 termos mais frequentes do campo selecionado, 'resumo' neste caso.

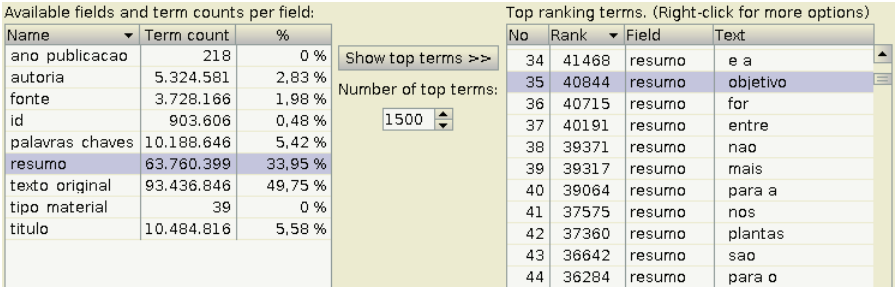

Figura 1. Tela do software Luke.

Geralmente, as *stop lists* são construídas em função de listas pré-existentes e disponibilizadas para aplicações gerais. Em situações em que as informações são muito dinâmicas, como na Internet, esta é uma boa abordagem. No entanto, em casos em que as informações são menos dinâmicas, como no Ainfo por exemplo, que contém a produção bibliográfica da Embrapa, as *stop lists* podem ser construídas a partir dos próprios termos presentes nos documentos catalogados. Desta forma, as *stop words* são diferentes de acordo com o domínio da aplicação. A natureza da informação armazenada e até mesmo as análises dos textos feitas durante

a indexação e a busca também devem ser consideradas na determinação das *stop words*. Por isso, os campos de informação são tão importantes. Cada campo pode estar associado a diferentes analisadores e tipos de dados e, portanto, ter uma stop list específica.

As Figuras 1 e 2 exibem dados indexados no Ainfo-Consulta. Enquanto a primeira enumera termos presentes nos resumos das obras, a segunda apresenta termos indexados no campo 'autoria', relacionado aos autores das obras. Nota-se que os termos relacionados nos dois casos são completamente diferentes, o que leva, portanto, a *stop lists* também distintas. Os analisadores utilizados também influenciam na seleção de *stop words*. Nos analisadores propostos por Marinho et al. (2012), por exemplo, os termos podem ser compostos por até cinco palavras adjacentes. Com isso, termos como 'e a', 'para a', 'para o' e 'm de', conforme ilustram as Figuras 1 e 2, são frequentes e devem ser considerados *stop words*, uma vez que não apresentam valor semântico. Esse tipo de stop word não é comum em listas pré-determinadas, mas pode ser detectado na abordagem aqui proposta, e explicada a seguir.

Em primeiro lugar, todo o conteúdo considerado é indexado sem a remoção de *stop words*. Depois que as listas de termos indexados são criadas para cada campo, avalia-se o valor semântico de cada termo para determinar quais devem ser consideradas *stop words* e compor a stop list. Como a lista de termos candidatos a *stop words* pode ser muito extensa - no exemplo considerado, são enumerados 186.827.317 termos - pode ser inviável analisá-la por completo. Então, inicia-se do mais frequente e prossegue-se até onde a disponibilidade de recursos humanos permitir. A permanência de termos com baixa frequência não gera impacto significativo. Finalmente, todo o conteúdo é indexado novamente excluindo-se as *stop words* selecionadas.

No	Rank	Field	Text
21	51669	autoria	o
22	42880	autoria	w
23	36817	autoria	oliveira
24	35882	autoria	i
25	32403	autoria	do
26	30998	autoria	santos
27	30064	autoria	dos
28	28834	autoria	filho
29	28790	autoria	k
30	28128	autoria	ed
31	25903	autoria	junior
32	24876	autoria	m de
33	24594	autoria	souza
34	23103	autoria	embrapa

Figura 2. Termos no campo 'autoria'.

A principal vantagem desta abordagem é a obtenção de melhor ranking dos resultados, uma vez que elimina a influência das *stop words* no cômputo da pontuação dos resultados. Em testes conduzidos no Ainfo-Consulta, não houve grande economia de espaço, mas tampouco efeitos negativos relevantes. A

metodologia aqui proposta, portanto, é uma alternativa interessante para a criação de stop lists que favoreçam a obtenção de melhores resultados em buscas realizadas em SRIs de domínio específico.

Referências

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Ainfo-Consulta**. Disponível em: <http://ainfo.cnptia.embrapa.br/consulta/>. Acesso em: 14 set. 2012.

MARINHO, I. J. P.; CARDONE, H. T. M.; VAZ, G. J. Evolução do mecanismo de busca do Ainfo-Consulta com uso de thesaurus agropecuário. In: CONGRESSO INTERINSTITUCIONAL DE INICIAÇÃO CIENTÍFICA, 6., 2012, Jaguariúna. **Anais**... Campinas: Embrapa: ITAL, 2012.

THE APACHE SOFTWARE FOUNDATION. **Solr Wiki**: language analysis. Disponível em: <http://wiki.apache.org/solr/LanguageAnalysis>. Acesso em: 25 jun. 2012.