Analisadores complementares para melhorar a qualidade das buscas em sistemas de recuperação de informação

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

Analisadores textuais que favorecem a obtenção de melhores resultados nas buscas em Sistemas de Recuperação de Informação (SRI) foram propostos por Marinho et al. (2012). Explorando as relações de equivalência presentes em tesauro, os autores conseguiram obter maior cobertura, métrica relacionada à capacidade em se recuperar os documentos que são relevantes para o usuário. O presente trabalho complementa essa solução com a proposta de um analisador adicional para obter melhor ranqueamento dos resultados em buscas que envolvam sinônimos, ou ainda, palavras diferenciadas por acentos. O recurso de autocompletar, que auxilia o usuário com sugestões de termos à medida que os caracteres são digitados no campo de busca, também é viabilizado por esse analisador complementar. Os resultados foram verificados no SRI Ainfo-Consulta, que possibilita consultas à produção bibliográfica da Embrapa (EMBRAPA INFORMÁTICA AGROPECUÁRIA, 2012).

O Ainfo-Consulta é construído com a plataforma de busca Apache Solr. A análise de texto nessa tecnologia é feita com um analisador composto de um *tokenyzer*, que gera uma lista de *tokens* a partir de um fluxo de texto, e zero ou mais filtros, que modificam essa lista. A Figura 1(a) mostra o analisador de indexação proposto por Marinho et al. (2012), com os seguintes ajustes:

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

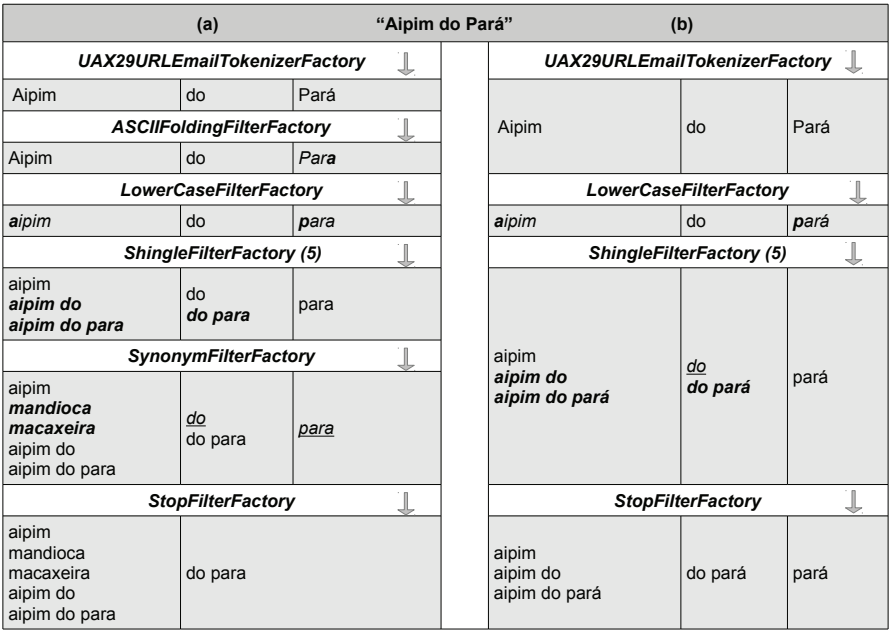

Figura 1. Analisadores: (a) proposto por Marinho et. al. (b) complementar.

- O *tokenizer UAX29URLEmailTokenizerFactory* substituiu o *StandardTokenizerFactory* para também reconhecer e classificar URLs e endereços de e-mail e de IP.
- O filtro *ISOLatin1AccentFilterFactory* foi substituído pelo *ASCIIFoldingFilterFactory*, mais atualizado e abrangente do que o primeiro. Esses filtros removem acentos das palavras a fim de tornar a busca mais simples e tolerante a erros de acentuação.
- O parâmetro do *ShingleFilterFactory*, filtro que cria termos compostos por mais de uma palavra, foi alterado de 3 para 5 a fim de tornar possível a exploração de sinônimos de termos compostos por até 5 palavras, situação que ocorre, por exemplo, no Thesagro, tesauro brasileiro especializado em literatura agrícola. Além disso, as sugestões de autocompletar também podem ser formadas por termos com até 5 palavras.

Os demais filtros utilizados são o *LowerCaseFilterFactory*, que substitui as letras maiúsculas por minúsculas, o *SynonymFilterFactory*, que acrescenta termos sinônimos em uma mesma posição, e o *StopFilterFactory*, que

elimina palavras irrelevantes (*stop words*). A ordem em que esses filtros são aplicados é fundamental para o adequado funcionamento do SRI e foi detalhadamente explicada por Marinho et. al. (2012).

A Figura 1(b) apresenta um analisador complementar que usa o *tokenizer UAX29URLEmailTokenizerFactory* e os filtros *LowerCaseFilterFactory*, *ShingleFilterFactory* e *StopFilterFactory*. Na Figura 1, as análises são ilustradas com a frase 'Aipim do Pará'. O filtro *ASCIIFoldingFilterFactory* não é utilizado no analisador complementar para que termos acentuados possam ser indexados e buscados. No primeiro analisador do exemplo, o termo 'Pará' transforma-se em 'para'. Como esse termo representa uma stop word devido à preposição 'para', não é indexado ou buscado isoladamente. Portanto, uma busca ao termo 'Pará', referente ao estado brasileiro, só é viabilizado com um analisador que mantém acentos.

Em relação ao ranqueamento, é desejável que uma busca ao termo 'aipim', por exemplo, retorne, em primeiro lugar, documentos que apresentem esse termo, em relação àqueles que contêm apenas os termos sinônimos. A Figura 1 mostra que enquanto 'aipim' consta no resultado do processamento dos dois analisadores, seus sinônimos aparecem em apenas um. Isso ocorre porque o filtro *SynonymFilterFactory* não compõe o analisador complementar. Assim, ajustando-se adequadamente os valores de relevância correspondentes aos dois analisadores, obtém-se o efeito esperado no ranking.

Finalmente, para o recurso de autocompletar, apenas o analisador complementar é utilizado. A Figura 2 ilustra as sugestões oferecidas pelo sistema após a digitação de 'consumo d' no campo de busca. O usuário obtém sugestões de termos que estão presentes nos documentos indexados, ordenados por suas frequências no índice.

Figura 2. Auto-completar.

O resultado da indexação realizada com o analisador complementar ocupa muito espaço. No Ainfo-Consulta, por exemplo,

equivale à cerca de 50% do índice. Ainda assim, desde que haja recursos suficientes, recomenda-se a combinação dos dois analisadores descritos neste trabalho como padrão para campos de conteúdos textuais, devido à viabilização do recurso de autocompletar e à maior qualidade alcançada em buscas envolvendo sinônimos ou palavras diferenciadas por acentos.

Referências

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Ainfo**. Disponível em: <http://www.ainfo.cnptia.embrapa.br>. 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. p. 1-9. CIIC 2012. No 12610.