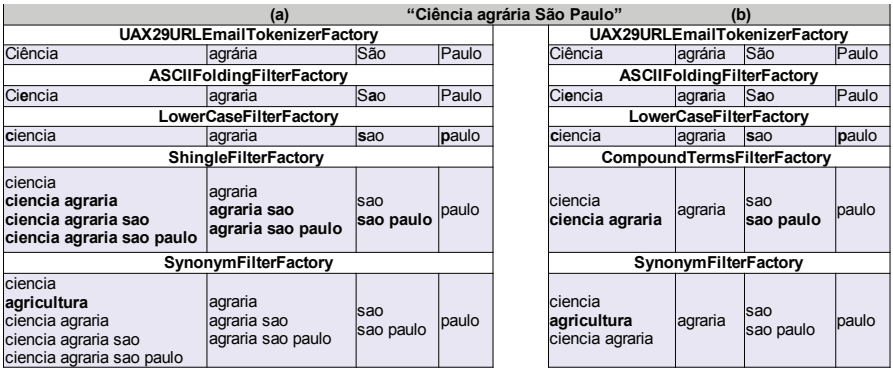
Tratamento eficiente de termos compostos na tecnologia Solr

Diego Felipe Zanardo[1]
Glauber José Vaz[2]

Um mecanismo de busca envolve análises de texto tanto na fase em que se processam os documentos para indexar seus termos quanto na fase de busca, que ocorre após o usuário entrar com uma consulta. O tratamento adequado de termos compostos é fundamental em sistemas mais avançados, uma vez que é muito comum a necessidade de indexar não somente palavras isoladas, mas também combinações delas que tenham um significado particular. Inúmeras localizações geográficas, por exemplo, são identificadas por termos compostos. A tecnologia Apache Solr, amplamente utilizada na construção de mecanismos de busca, disponibiliza um componente de software, *ShingleFilterFactory*, que forma termos compostos a partir de palavras em sequência, mas não é muito eficiente. Neste trabalho, desenvolvemos um componente de software, *CompoundTermsFilterFactory*, que faz o tratamento de termos compostos de maneira mais eficiente usando uma lista de termos compostos considerados relevantes, como aqueles que indicam uma localização geográfica ou que apresentam sinônimos.

A análise de texto na tecnologia Apache Solr é feita a partir de um analisador formado por um *tokenizer*, que divide um fluxo de texto em *tokens*, e zero ou mais filtros, que alteram esses *tokens*. A Figura 1 mostra, para a entrada 'Ciência agrária São Paulo', o funcionamento de dois analisadores similares. Tanto o analisador da Figura 1a quanto o da Figura 1b utilizam o *tokenizer UAX29URLEmailTokenizerFactory*, que separa o texto em *tokens*, reconhecendo e classificando URLs e endereços de e-mails e de IP. Ambos também utilizam os filtros *ASCIIFoldingFilterFactory*, que remove acentos das palavras, *LowerCaseFilterFactory*, que substitui as

[1] Fatec/Americana - diegozanardo@yahoo.com.br
[2] Embrapa Informática Agropecuária - glauber.vaz@embrapa.br



(a)	“Ciência agrária São Paulo”			(b)			
UAX29URLEmailTokenizerFactory				**UAX29URLEmailTokenizerFactory**			
Ciência	agrária	São	Paulo	Ciência	agrária	São	Paulo
ASCIIFoldingFilterFactory				**ASCIIFoldingFilterFactory**			
Ciencia	agraria	Sao	Paulo	Ciencia	agraria	Sao	Paulo
LowerCaseFilterFactory				**LowerCaseFilterFactory**			
ciencia	agraria	sao	paulo	ciencia	agraria	sao	paulo
ShingleFilterFactory				**CompoundTermsFilterFactory**			
ciencia **ciencia agraria** **ciencia agraria sao** **ciencia agraria sao paulo**	agraria **agraria sao** **agraria sao paulo**	sao **sao paulo**	paulo	ciencia **ciencia agraria**	agraria	sao **sao paulo**	paulo
SynonymFilterFactory				**SynonymFilterFactory**			
ciencia **agricultura** ciencia agraria ciencia agraria sao ciencia agraria sao paulo	agraria agraria sao agraria sao paulo	sao sao paulo	paulo	ciencia **agricultura** ciencia agraria	agraria	sao sao paulo	paulo

Figura 1. Analisadores utilizando os filtros (a)*ShingleFilterFactory* e (b)*CompoundTermsFilterFactory*.

letras maiúsculas por minúsculas, e *SynonymFilterFactory*, que acrescenta termos sinônimos. Porém, para o tratamento de termos compostos, usam filtros diferentes: o *ShingleFilterFactory*, já disponível na tecnologia Solr, e o *CompoundTermsFilterFactory*, desenvolvido neste trabalho. Marinho et al. (2012) explicam detalhadamente analisadores parecidos com esses.

O filtro *ShingleFilterFactory*, conforme pode ser observado na Figura 1a, forma todos os termos compostos possíveis para a sequência de texto de entrada, limitado apenas pela quantidade máxima de palavras, preestabelecida por parâmetro. Então, se este parâmetro fosse estabelecido em três, por exemplo, ‘ciencia agraria sao paulo’, não seria adicionado na análise por conter quatro palavras. Já o *CompoundTermsFilterFactory* forma termos compostos a partir de palavras em sequência somente se o candidato a termo constar na lista de termos compostos relevantes. No exemplo exposto, ‘ciencia agraria’ e ‘sao paulo’ fazem parte dessa lista. Sem filtros como esses, não seria possível, por exemplo, tratar sinônimos adequadamente, uma vez que ‘agricultura’ e ‘ciencia agraria’ não seriam identificados como equivalentes pelo fato de ‘ciencia’ e ‘agraria’ serem manipulados estritamente como termos isolados. De maneira análoga, ‘sao paulo’ só é reconhecido como um termo único devido aos filtros em questão. Isso possibilita, por exemplo, a exibição de um mapa quando se trata de localização geográfica.

Embora ambos os filtros possibilitem a identificação de termos compostos, o *CompoundTermsFilterFactory* representa um grande avanço em relação ao *ShingleFilterFactory*, pois forma apenas termos compostos considerados relevantes, ou seja, aqueles que estão presentes em uma lista predeterminada. Com isso, utiliza muito menos espaço em memória e tempo de processamento.

Para efeito de comparação entre os filtros, utilizamos os dois analisadores representados na Figura 1 para indexar a base de dados do Ainfo (EMBRAPA INFORMÁTICA AGROPECUÁRIA, 2013), que conta com cerca de 900 mil registros dos acervos impressos e digitais da Embrapa. O *ShingleFilterFactory* foi configurado para formar termos com até cinco palavras, enquanto o *CompoundTermsFilterFactory* considerou termos compostos relevantes todos aqueles descritores que constam no Thesagro, thesaurus brasileiro especializado em literatura agrícola (BRASIL, 1999). Quando utilizamos o *CompoundTermsFilterFactory*, menos de nove minutos foram suficientes para construir os índices, que ocuparam cerca de 1.1 GB de espaço em disco. No caso do *ShingleFilterFactory*, mais de 40 minutos e cerca de 8.1 GB foram necessários para construir os índices.

Portanto, a utilização do filtro *CompoundTermsFilterFactory*, desenvolvido neste trabalho, possibilita um tratamento muito mais eficiente de termos compostos nos mecanismos de busca do que o ShingleFilterFactory, tanto no que se refere a espaço em disco quanto em tempo de processamento para indexação.

Referências

BRASIL. Ministério da Agricultura e do Abastecimento. Biblioteca Nacional de Agricultura. **Thesagro**: Thesaurus Agrícola Nacional. Brasília, DF, 1999.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Ainfo**. [2013]. Disponível em: <http://www.ainfo.cnptia.embrapa.br/>. Acesso em: 26 set. 2013.

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.