Análise temporal de tópicos de publicações em agroinformática

Caio de Sá Lopes[1]
Ricardo Marcondes Marcacini[1]
Solange Oliveira Rezende[1]
Maria Fernanda Moura[2]

O avanço das tecnologias para aquisição e armazenamento de dados tem permitido que o volume de informação gerado em formato digital aumente de forma significativa nas organizações. Cerca de 80% desses dados estão em formato não estruturado, no qual uma parte significativa são textos. A organização inteligente dessas coleções textuais é de grande interesse para a maioria das instituições, pois agiliza processos de busca, recuperação e análise da informação. Nesse contexto, a Mineração de Textos permite a transformação desse grande volume de dados textuais não estruturados em conhecimento útil, muitas vezes inovador para as organizações (REZENDE et al., 2011).

A análise exploratória de informações publicadas na web é uma tarefa relevante para diversas aplicações. Em especial, o monitoramento de tópicos emergentes a partir de textos tem recebido grande atenção na literatura, como a análise de tendências de tópicos extraídos de notícias, artigos científicos e redes sociais. Um exemplo desse tipo de análise pode ser visto na Figura 1, com o uso da ferramenta Torch-ETS (PANNAGIO et al., 2011). Neste figura, é ilustrado um agrupamento hierárquico (Figura 1A), os tópicos existentes na coleção de texto (Figura 1B), a evolução temporal de um tópico selecionado pelo usuário (Figura 1C) e, por fim, os documentos relacionados ao tópico (Figura 1D).

O foco desse trabalho está no desenvolvimento de uma *Application Programming Interface* (API), para lidar com a leitura dos resultados em diversos padrões dos resultados que existem atualmente no projeto

[1] Universidade de São Paulo - caio.lopes@colaborador.embrapa.br, {rmm, solange}@icmc.usp.br

[2] Embrapa Informática Agropecuária - fernanda@cnptia.embrapa.br

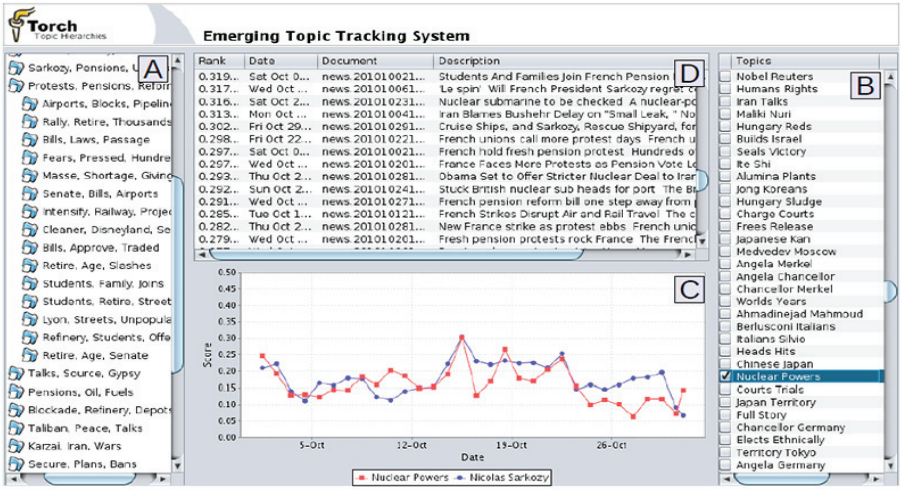

Figura 1. Resultado de uma análise temporal pela ferramenta Torch-ETS.

CRITIC@. Nesse projeto, várias ferramentas utilizadas em seus processos de Mineração de Textos geram diferentes formatos de resultados, todos especificados em algum padrão XML. Esses resultados correspondem a agrupamentos de textos, com tópicos descritos ou não, informação temporal sobre os tópicos, palavras-chaves, etc. Assim, todos os resultados precisam ser lidos e reutilizados em alguma parte do processo; tarefa que será viabilizada pela API em desenvolvimento.

Para a primeira versão da API, está-se trabalhando sobre o resultado do agrupamento gerado pela ferramenta Torch.

No entanto, a API será facilmente adaptada para ler os mais diversos padrões gerados. Com ela é possível recuperar, de forma transparente ao usuário, todas as informações disponíveis no arquivo XML que representa a hierarquia de tópicos, como descritores, série temporal, o tópico 'pai' e os tópicos 'filhos'. Para seu desenvolvimento, está sendo usada a linguagem de programação Java, com o *Simple API for XML* (SAX).

Por exemplo, na Figura 2 é mostrada uma base de dados já organizada em forma de hierarquia de tópicos, utilizando a API desenvolvida nesse trabalho.

A API atualmente se encontra em fase de desenvolvimento final e testes. Para uma próxima versão pretende-se estender as funcionalidades para grande coleções de texto.

Figura 2. Visualização da leitura feita pela API.

Referências

PANAGGIO, B. Z.; MARCACINI, R. M.; REZENDE, S. O. “Torch-ETS: análise exploratória de tópicos emergentes com apoio de agrupamento hierárquico de textos.” In: SIMPÓSIO BRASILEIRO DE SISTEMAS MULTIMÍDIA E WEB 17.; WORKSHOP DE FERRAMENTAS E APLICAÇÕES, 10., 2011, Florianópolis. **Anais**.... Porto Alegre: SBC, 2011. p. 143-147. Webmedia. WFA’2011.

REZENDE, S. O.; MARCACINI, R. M.; MOURA, M. F., “O uso da Mineração de Textos para Extração e Organização não Supervisionada de Conhecimento”. **Revista de Sistemas de Informação da FSMA**, n. 7, p. 7-21, 2011.