Compilação e Análise da Informação Utilizada ou Produzida pela Rede AgroHidro

Maria Fernanda Moura; Renan Gomes Pereira; Gabriel Mari Tararam; Luis Eduardo Gonzales; Celina Maki Takemura; Stanley Robson de Medeiros Oliveira; Silvio Roberto Medeiros Evangelista; Solange Oliveira Rezende; Fabiano Fernandes dos Santos

Resumo

Neste trabalho é apresentada a versão 2.0 do software CRITIC – Compilação e Recuperação de Informação Técnico-científica e Indução ao Conhecimento, com base em técnicas de mineração de textos sobre artigos científicos. O software usa um provedor de serviços para acesso aos repositórios de referências aos artigos, cujo acesso é aberto. Na sua versão atual, permite-se a realização de análise exploratória sobre os resultados das consultas, na qual são automaticamente identificados: tópicos hierárquicos dos temas cobertos na consulta; a distribuição temporal destes temas; e, a distribuição geoespacial dos temas cobertos pelos textos. O software CRITIC foi projetado para ser facilmente estendido e disponibilizado em repositório Open Source – a fim de possibilitar um desenvolvimento colaborativo de novas funcionalidades.

Termos para indexação: recuperação de informação, análise exploratória de dados, mineração de textos.

Compilation and Analysis of the Information Used or Produced by AgroHidro

Abstract

This paper presents the second version of the software CRITIC – compilation and retrieval of technical and scientific information and knowledge induction, based on text mining techniques on scientific papers. The software uses a service provider to access the repositories of references to articles, which are open access. In its current version, it provides an exploratory analysis of the search results, which automatically identify: hierarchical topics; temporal distribution of a specific topic; and, geospatial distribution of topics. CRITIC was designed to be easily expanded and placed in a Open Source repository, in order to have a collaborative development of new features.

Index terms: information retrieval, exploratory data analysis, text mining.

Introdução

A proposta do projeto *Compilação e Recuperação de Informação Técnico-científica e Indução ao Conhecimento de forma ágil na rede AgroHidro* (CRITIC@) é concentrar as ações de organização, compilação e análise da informação utilizada e produzida pela rede AgroHidro. Dessa forma, no projeto CRITIC@, são integradas soluções de tecnologia da informação que permitem:

1) Recuperação da informação de interesse da Rede AgroHidro a partir de provedores de publicações de acesso aberto, dos quais também fazem parte os metadados das produções científicas e tecnológicas da Embrapa.

2) Indexação de um repositório cópia dos provedores de interesse de acordo com vocabulário específico de uso da rede.

3) Busca simples e avançada compatíveis com a ferramenta Sistema Aberto e Integrado de Informação em Agricultura (SABIIA) (VACARI et al., 2011) da Embrapa.

4) Apresentação de estatísticas básicas sobre a consulta, tais como, frequências em provedores, por autores, por idioma, por assuntos, ano de publicação, etc.

5) Utilização de filtro simples de busca sobre cada item descrito pelas estatísticas básicas, por exemplo, refazer a busca apenas para determinados assuntos e autores.

6) Realização de uma análise exploratória mais detalhada dos resultados de busca, para: (a) exibir os documentos por grupos hierárquicos de assuntos automaticamente identificados; (b) exibir os documentos de um grupo em um mapa, de acordo com sua distribuição geoespacial; (c) exibir os assuntos identificados em um grupo de acordo com sua distribuição temporal; e (d) filtrar assuntos e intervalos de tempo para serem graficamente exibidos. Com seu uso, pode-se, por exemplo, identificar quais culturas são praticadas em quais regiões ou em quais bacias hidrográficas do país.

A primeira versão do CRITIC (MOURA et al., 2015), software resultante do projeto CRITIC@, tratava-se de um beta-teste desenvolvido sobre a ferramenta SABIIA, pois esta contempla os repositórios de interesse da Rede AgroHidro, é amplamente utilizada e já provia o arsenal de armazenamento e busca necessários. Na versão beta-teste, o interesse era avaliar as técnicas de análise de dados apresentadas à rede e formas de visualização desses resultados. Na versão 2.0 do CRITIC, houve a troca de toda a arquitetura do software: (a) substituição dos processos anteriormente executados pela SABIIA, tais como a persistência dos dados, indexação, busca e visualização dos resultados de busca; (b) atualização da indexação e pré-processamento dos textos, de forma incremental e paralela; (c) troca do processo de geoespacialização dos documentos; (d) processo de busca; e, (e) componentes de visualização dos resultados da análise exploratória dos resultados de busca. Os processos são revistos no item material e métodos, desde a persistência de dados até os processos de análise implementados. A seguir, um resultado de busca e da análise exploratória é discutido, bem como possíveis desmembramentos desse ferramental.

Material e Métodos

Os processos implementados no ambiente de software CRITIC 2.0 são:

1) Atualização incremental do repositório de documentos: os documentos são recuperados via JOAI harvester (JOAI SOFTWARE, 2016), incrementalmente de mês em mês a partir de uma lista de provedores de interesse. As pastas de documentos com as atualizações são submetidas ao pré-processamento e depois juntadas às pastas anteriores, via Shell scripts.
2) Pré-processamento: composto por duas etapas (a) geoespacialização e (b) indexação. Para a geoespacialização extraem-se entidades nomeadas, com o uso de uma ferramenta linguística de nome OpenCalais (THOMSON REUTERS, 2016). Como o OpenCalais trabalha apenas com textos em inglês, utiliza-se um processo de tradução para os textos em outras línguas. Identificados os topônimos, há um processo de desambiguação para georeferenciar cada texto; então,

essa informação é incluída nos metadados do texto. O processo de indexação é incremental, isto é, indexa apenas os novos textos considerando a indexação realizada nas atualizações anteriores e, pode ser paralelizado. A indexação é realizada com base em um vocabulário fixado para a Rede AgroHidro (UNGARO; MOURA, 2013), que pode ser reconfigurado. O filtro de vocabulário é aplicado apenas aos campos título, resumo, descrição, palavras-chaves e texto completo (quando existente).

3) Processo de busca: a nova busca é semelhante àquela existente na ferramenta SABIIA, sendo que as expressões podem especificar campos específicos dos documentos (título, autor, etc), usar operadores booleanos e distância de caracteres para termos lexicamente semelhantes. As estatísticas mostradas na apresentação de resultados, tais como, número de provedores, etc, foram mantidas, bem como novos filtros de busca a partir dos resultados de busca (por exemplo, só tais autores e (ou) um intervalo de tempo).

4) Análise exploratória dos resultados de busca: os resultados são divididos em: (a) tópicos hierárquicos, obtidos por agrupamento hierárquico de documentos e descrição dos grupos; para cada tópico há um gráfico de sua (b) cobertura temporal dentre os resultados de busca, considerando a distribuição conjunta dos descritores do tópico; e, de sua (c) cobertura geográfica – retângulos envolventes de cada texto do tópico.

5) Visualização da análise exploratória: foram utilizados componentes da biblioteca Java D3 (D3.JS , 2016) para as hierarquias e componente Google para mapas.

Resultados e Discussão

Seja a expressão de busca "(mata atlantica)OR(polinizacao)OR(biomassa)", cujos resultados e análise exploratória de um dos tópicos são apresentados na Figura 1. A parte Busca da Figura 1 tem interface compatível com as ferramentas de busca de documentos da Embrapa, já amplamente difundidas. Ainda, na Busca, à esquerda, é possível observar as estatísticas descritivas de cada assunto de interesse, sejam palavras-cha-

ves, nomes dos autores, repositórios, tipos de publicação, etc; e, nessa mesma parte da tela, é possível realizar mais filtros sobre os resultados de busca. Acionando o botão Analisar Busca, são apresentados os tópicos hierárquicos estimados para aquele resultado de busca, como mostrado na parte Tópicos da Figura 1; e, na parte Tópicos TreeMap, uma forma diferenciada de se observar a mesma hierarquia. A cada tópico corresponde: (a) um mapa com a cobertura geográfica do mesmo – todos os polígonos envolventes dos textos do tópico; (b) o conjunto de textos do tópico; e (c) gráfico com a série temporal correspondente ao tópico dentro do resultado de busca. Na Figura 1, foi ilustrado o tópico referente a "crescimento de espécies e biomassa na área de floresta da Mata Atlântica", cujos descritores encontrados foram "Especies, Atlantica, Crescimento, Area, Floresta, Biomassa". Essa análise de dados também pode ser filtrada para que se possa ir delimitando assuntos mais específicos.

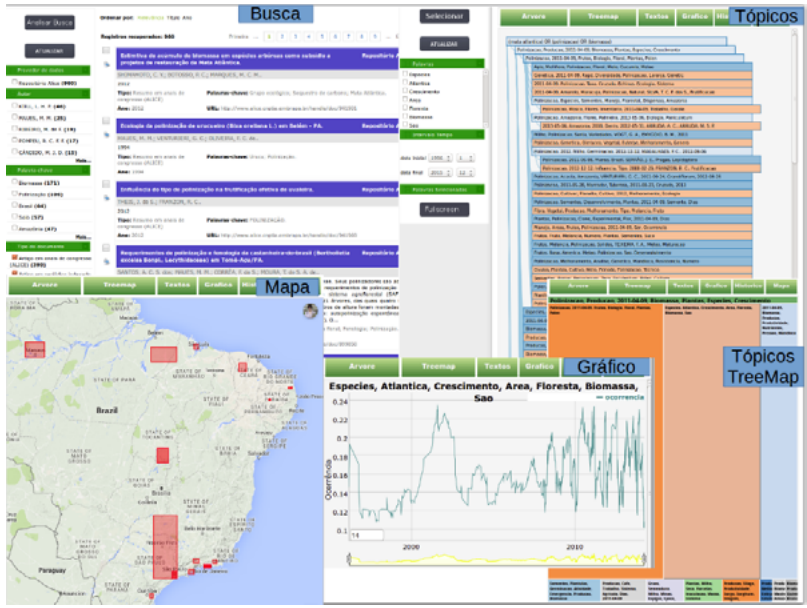

Figura 1. Exemplo de busca e exploração de seus resultados.

Conclusões

Com a versão 2.0 do software CRITIC, a partir dos dados de interesse da Rede Agrohidro, em repositórios de dados Open Access, pode-se: especificar com precisão uma expressão de busca; filtrar os resultados dessa busca; explorar os resultados dessa busca por meio de tópicos hierárquicos inferidos; e, para cada tópico, explorar sua distribuição geográfica e temporal.

O software foi desenvolvido para ser Open Source, facilmente reconfigurável e expandido; a fim que se possa ter um desenvolvimento colaborativo de novas funcionalidades. Após os testes finais de estabilidade, ele deverá ser colocado em um repositório público. Futuramente, ele pode ser integrado a outros softwares de busca da Embrapa, ou de quaisquer outras instituições que tenham interesse no mesmo.

Referências

D3.JS - DATA-driven documents. Disponível em: <https://d3js.org/>. Acesso em: 20 maio 2016.

JOAI SOFTWARE digital library for earth system education. Disponível em: <http://www.dlese.org/dds/services/joai_software.jsp>. Acesso em: 20 maio 2016.

MOURA, M. F.; TARARAM, G. M.; MARCACINI, R. M.; GONZALES, L. E; TAKEMURA, C. M.; SILVA, L. E. A.; SANTOS, F. F. dos; REZENDE, S. O.; EVANGELISTA, S. R. M. Um software para recuperar e analisar artigos Open Access em agricultura utilizando técnicas de mineração de textos. In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 10., 2015, Ponta Grossa. **Uso de VANTs e sensores para avanços no agronegócio**: anais. Ponta Grossa: Universidade Estadual de Ponta Grossa, 2015.

THOMSON REUTERS. **OpenCalais**. Disponível em: <http://www.opencalais.com/>. Acesso em: 9 maio 2016.

UNGARO, F. P.; MOURA, M. F. Análise de tendências tecnológicas em recursos hídricos. In: MOSTRA DE ESTAGIÁRIOS E BOLSISTAS DA EMBRAPA INFORMÁTICA AGROPECUÁRIA, 9., 2013, Campinas. **Resumos**... Brasília, DF: Embrapa, 2013. p. 82-85.

VACARI, I.; VISOLI, M. C.; GONZALES, L. E. Acesso aberto a informação científica agropecuária na internet: caso do sistema aberto e integrado de informação em agricultura (Sabiia). In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA, 8., 2011, Bento Gonçalves. **Anais**... Florianópolis: UFSC; Pelotas: UFPel, 2011.