Componentes para a integração e extração de padrões em textos para versão 1.0 do Ambiente CRITIC@

Leandro Eduardo Annibal Silva[1]
Maria Fernanda Moura[2]

O projeto Compilação e Recuperação de Informações Técnico-científicas e Indução ao Conhecimento de forma Ágil na Rede AgroHidro(CRITIC@) consiste em melhorar a gestão do conhecimento técnico-científico na área de recursos hídricos, por meio de análises cruzadas das informações, bem como subsidiar ações de investigação e disseminação do conhecimento na rede de pesquisa. Esse projeto tem como objetivos: permitir busca e visualização de informações sobre recursos hídricos, de maneira eficiente; permitir, por meio do ambiente semiautomático da metodologia, realizar análises de tendências técnico-científicas na área a partir da informação organizada; construir e validar uma metodologia semiautomática de organização e análise da informação técnico-científica do domínio de conhecimento de recursos hídricos coberto pela rede AgroHidro; auxiliar *screening* tecnológico; gerar um glossário técnico-científico da área de recursos hídricos, para auxílio à organização da informação, sua catalogação e seu armazenamento, busca e análise; obter uma representação ontológica do conhecimento técnico-científico detido pela Rede AgroHidro. Por ser complexo, o projeto foi dividido em três etapas:

1) Constrói-se, ainda em 2014, a metodologia de organização e análise, independentemente de ter ou não conhecimento organizado.
2) Evolui-se, em 2015, a metodologia de organização e análise para utilizar *thesaurus* (e/ou uma relação taxonômica).

[1] Pontifícia Universidade Católica de Campinas
[2] Embrapa Informática Agropecuária

3) Evolui-se, em 2016, a metodologia de organização e análise para utilizar uma ontologia. Atualmente, estando na primeira etapa, temos como meta desenvolver uma metodologia de organização e análise utilizando extração estatística de informação.

Nesta primeira etapa, uma decisão de projeto foi desenvolver programas independentes, que se comunicam via arquivo de dados. Desta forma, ganhou-se em flexibilidade, pois os programas podem ser executados via linha de comando, podem ser integrados via ferramentas de *workflow* ou serem ativados por outros programas, bem como desenvolvidos em diferentes linguagens e padrões de desenvolvimento. Outra decisão foi utilizar o Sistema Aberto e Integrado de Informação em Agricultura (SABIIA) (HIRATA; VACARI, 2010) como ferramenta de busca, na primeira etapa do CRITIC@, porque este contempla os dados produzidos e utilizados pela rede AgroHidro. Além disso, o SABIIA faz colheita de dados no padrão *Open Archive Initiative* (OAI) (OPEN ARCHIVES INITIATIVE, 2014), em vários repositórios de artigos e materiais científicos de acesso aberto. Conforme ilustrado na Figura 1, na parte mensal, temos o Banco de Dados do SABIIA,

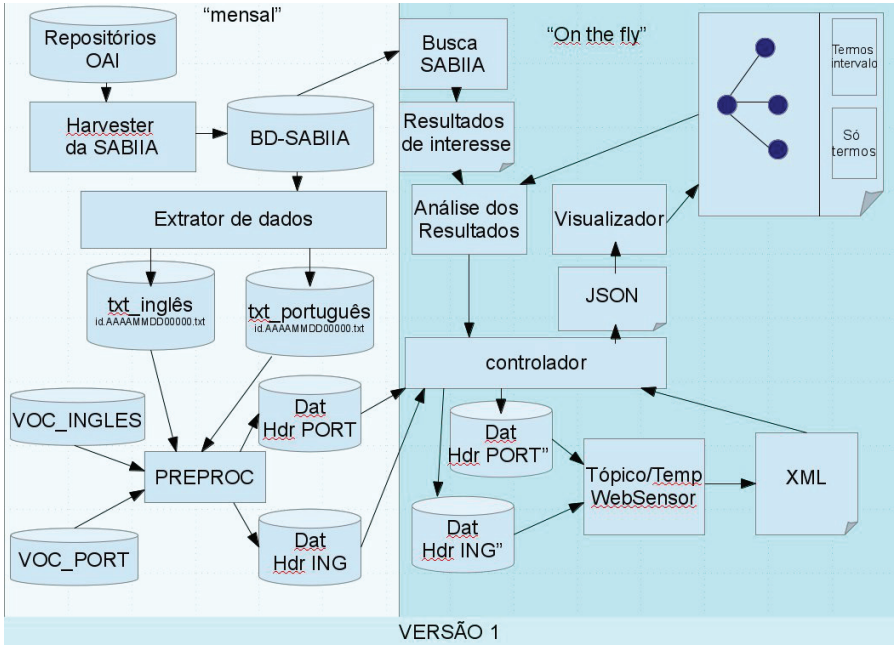

Figura 1. Versão 1 do Ambiente CRITIC@.

onde estão arquivados todos os artigos, notícias, etc, de interesse da rede AgroHidro. O Extrator de dados é um programa que a partir do BD-SABIIA, recupera todos os dados armazenados e os salva em aquivos TXT. Os arquivos TXT ficam na máquina local para que possam ser realizados os processos de pré-processamento; além disso, são separados por idioma (português e inglês). Após a separação, aplica-se o processo PREPROC, para transformar a representação textual em uma matricial. O PREPROC utiliza uma lista de vocábulos previamente fixados e presentes nos textos analisados como colunas das matrizes. As linhas dessas matrizes correspondem a cada documento; e, cada célula contém a frequência de ocorrência do vocábulo no texto. São geradas duas matrizes (arquivos DAT e HDR), uma com os textos em inglês e outra com os textos em português. Na parte "On the Fly", que é utilizada pelo cliente, a partir de uma busca no SABIIA é gerado um arquivo com os IDs recuperados para que o Controlador possa processar os dados: a) criar sub arquivos DAT e HDR correspondentes aos Ids; b) gerar um xml com tópicos e informação temporal a partir da chamada à ferramenta WebSensor; c) a partir desse xml o Controlador gera um arquivo JSON para uso do Visualizador. O Visualizador usa JAVA script junto à biblioteca D3 Data Driven Documents (2014), formando um arquivo html interpretável pelo *browser*, o qual apresenta os resultados ao usuário.

Neste plano de trabalho foram desenvolvidos o Extrator, o Controlador e o Visualizador. As demais ferramentas foram configuradas para serem integradas ao sistema. O próximo passo é integrar a execução com os resultados do SABIIA, via web, para disponibilizar o uso das ferramentas ao usuário. Desta forma, este trabalho contribui para o desenvolvimento da etapa 1 do CRITIC@, com estes componentes.

Palavras-chave: componentes de software, reúso, mineração de textos.

Referências

D3 DATA Driven Documents. 2014. Disponível em: <http://d3js.org/>. Acesso em: 25 set. 2014.

HIRATA, A.; VACARI, I. Uso de software livre para implementação de provedores de serviços OAI-PMH: caso do provedor de serviços Sabiia. In: MOSTRA DE ESTAGIÁRIOS E BOLSISTAS DA EMBRAPA INFORMÁTICA AGROPECUÁRIA, 6., 2010, Campinas. **Resumos**... Campinas, 2010. p. 10-14.

OPEN ARCHIVES Initiative. Disponível em: <http://www.openarchives.org/>. Acesso em: 25 set. 2014.