Desenvolvimento de um ambiente facilitador de integração de ferramentas de mineração de textos

Cesar Haruaki Takagi[1]
Fabiano Fernandes dos Santos[2]
Solange Oliveira Rezende[2]
Maria Fernanda Moura[3]

O objetivo deste trabalho é criar um ambiente que integre as várias ferramentas desenvolvidas e utilizadas nos projetos Tiena (MOURA et al., 2010) e Ainfo6 (PRAXEDES et al., 2009).

De acordo com Santos (2010), mineração de textos é um conjunto de técnicas e processos que descobre conhecimento inovador nos textos. Atualmente ela é bastante utilizada para extrair conhecimento de grandes coleções de documentos textuais, que podem ser úteis para uma tomada de decisão. O processo de mineração de textos pode ser dividido em cinco grandes fases: Identificação do Problema, Pré-processamento, Extração de Padrões, Pós-processamento e Utilização do Conhecimento. E, por ser um processo complexo, é necessária a utilização de ferramentas específicas para cada fase da mineração. Ainda, atualmente conta-se com uma boa disponibilidade de ferramentas de domínio público, que podem ser utilizadas conforme as necessidades específicas de cada aplicação em mineração de textos.

Nos projetos Tiena e Ainfo vem sendo utilizada a ferramenta TaxEdit - Taxonomy Editor (MOURA at al., 2011) para realizar a integração das diversas etapas de mineração de textos. A TaxEdit mantém as ferramentas

[1] USP, São Carlos, estagiário da área Inteligência Computacional, chtakagi@gmail.com
[2] ICMC USP São Carlos, {fabianof, solange}@icmc.usp.br
[3] Embrapa Informática Agropecuária, maria-fernanda.moura@embrapa.br

de cada etapa altamente acopladas, dificultando a troca e teste de outras ferramentas. Um novo ambiente que possibilitasse a integração das ferramentas de mineração solucionaria esse problema.

Para o desenvolvimento do ambiente, foi decidida a utilização do conceito de *workflow*, definido como "automação do processo de negócio, na sua totalidade ou em partes, onde documentos, informações ou tarefas são passadas de um participante para o outro para execução de uma ação, de acordo com um conjunto de regras de procedimentos" (WfMC, 2012). Esse conceito foi adotado devido ao fato de ter sido observado que o processo de mineração de textos se assemelha a um *workflow*, em que os dados são processados em cada etapa e os seus resultados são passados para a etapa seguinte. Com isso foi definida a criação de um arquivo de *workflow*, onde seriam descritos os componentes, ou seja, as ferramentas a serem utilizadas em cada etapa do processo de mineração, bem como seus dados e parâmetros. Para a criação de arquivos de *workflow*, foram pesquisadas várias ferramentas de criação de workflows, e dentre elas foi escolhida a ferramenta Kepler (THE KEPLER PROJECT, 2012), devido a sua praticidade na criação e armazenamento de *workflows*, e pela interface altamente intuitiva.

Após essas considerações, foi desenvolvido o ambiente WorkFlow to Execution Engine (WF2E), cuja arquitetura encontra-se sumarizada na Figura 1. Esse ambiente, a partir de um arquivo de *workflow*, lê seus componentes e parâmetros, procura as suas ferramentas correspondentes e dados a serem processados, realiza o processamento e envia os resultados para o componente da etapa seguinte. Com isso, é possível combinar e testar as várias ferramentas de cada etapa da mineração de texto, bastando editar ou utilizar outro arquivo de *workflow*.

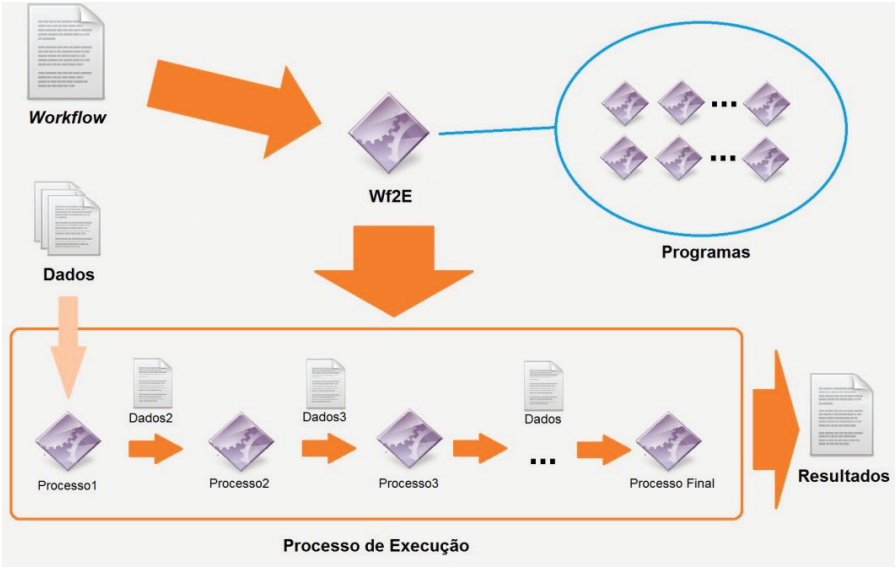

Figura 1. Arquitetura sumarizada da ferramenta Wf2E.

Agradecimentos

À equipe do Laboratório de Inteligência Computacional, do ICMC/USP, pelas colaborações sempre presentes. Aos pesquisadores Roberto Hiroshi Higa e Sérgio Aparecido Braga da Cruz (Embrapa Informática Agropecuária), pelas colaborações e ideias iniciais.

Referências

MOURA, M. F.; MERCANTI, E.; PEIXOTO, B. M.; MARCACINI, R. M.; TAMADA, T.; LIMA, A. F.; SANTOS, F. F. dos. **TaxEdit** - Taxonomy Editor V 2.0. Versão 1.0. Campinas: Embrapa Informática Agropecuária, 2011. 1 CD-ROM.

MOURA, M. F.; SANTOS, A. D. dos; JORGE, A. M. G.; SPERANZA, E. A.; ASSAD, E. D.; ESCUDEIRO, N. F. V.; SANTOS, F. F. dos; OLIVEIRA, L. H. M. de; CONRADO, M. da S.; HIGA, R. H.; ROSSI, R. G.; MARCACINI, R. M.; REZENDE, S. O. **TIENA** – Tecnologias Inovadoras em mineração de textos para apoio à Espacialização de Notícias Agrícolas. Campinas: Embrapa Informática Agropecuária, 2010. (Embrapa. Macroprograma 3) - Projeto - 03.10.01.02400.00. Projeto em Andamento.

PRAXEDES, M. G. G.; FARIA, A. L. D. de; ARRUDA, R. G.; CASTRO, R. L.; VACARI, I.; GAMA, G. F. de B.; SIMÃO, V. P. M. **Evolução do software Ainfo6 com uso de ferramentas da Web Semântica e mineração de textos e digitalização da produção científica**. Campinas: Embrapa Informática Agropecuária, 2009. 15 p. (Embrapa. Macroprograma 5 - Institucional. Projeto - 05.08.09.002.00.00. Projeto em andamento.

SANTOS, F. F. dos. **Selecionando candidatos a descritores para agrupamentos hierárquicos de documentos utilizando regras de associação**. 2010. 79 p. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional) Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos,SP. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-17112010-110417/>.

THE KEPLER Project. “**The Kepler Project** - Kepler” Disponível em: <https://kepler-project.org/>. Acesso em: 23 maio 2012.

WfMC. “**Workflow Management Coalition**”. Disponível em: <http://www.wfmc.org/>. Acesso em: 27 jun.2012.