Análise de tendências em pagamentos por serviços ambientais hídricos

Ivan Prado da Costa[1]
Azeneth Eufrazino Schuler[2]
Rachel Bardy Prado[2]
Maria Fernanda Moura[3]

O projeto Compilação e Recuperação de Informações Técnico-científicas e Indução ao Conhecimento de forma Ágil na Rede AgroHidro (CRÍTiC@) se propõe a concentrar as ações de análise e organização sistematizada da informação utilizada e produzida pelo projeto "Impactos do uso agrícola e das mudanças climáticas sobre os recursos hídricos em diferentes ecorregiões brasileiras: diagnose e estratégias mitigadoras". Seu objetivo é organizar e analisar a informação técnico-científica disponível na rede AgroHidro para apoio à gestão do conhecimento e da inovação. Baseado na proposta do projeto CRÍTiC@ e em seus objetivos, neste trabalho procura-se contribuir para que sejam cumpridas metas específicas, como viabilizar a extração semiautomática de palavras-chaves e tópicos em textos do domínio de recursos hídricos, bem como observar sua distribuição e relações, especialmente as hierárquicas – tópicos e sub-tópicos. Espera-se, por meio dos tópicos obtidos, encontrar semelhanças entre os arquivos da base de dados e identificar tendências.

Inicialmente foram realizados estudos de ferramentas para mineração de textos, bem como seus respectivos métodos estatísticos de extração de informação e padrões, estatísticas descritivas dos tópicos encontrados e suas visualizações em hierarquias e gráficos. Além disso, durante a realização dos experimentos, foi possível avaliar a inclusão de termos a arquivos

[1] Universidade Estadual de Campinas - ivan.costa@colaborador.embrapa.br
[2] Embrapa Solos - {azeneth.schuler, rachel.prado}@embrapa.br
[3] Embrapa Informática Agropecuária - maria-fernanda.moura@embrapa.br

de vocabulário controlado gerando assim uma primeira versão de lista de termos do domínio.

Para a realização dos primeiros experimentos foram utilizados textos de pagamentos por serviços ambientais hídricos, em inglês e português. No total, foram selecionados 150 textos, sendo 117 em inglês e 33 em português. Esse problema foi identificado no projeto “Fortalecimento do conhecimento, organização da informação e elaboração de instrumentos de apoio aos Programas de Pagamentos por Serviços Ambientais Hídricos no meio rural” (do Macroprograma 5, sob a Rede AgroHidro), em sua atividade de “Organização e análise de informações textuais secundárias para subsídio a instrumentos de apoio aos programas de PSA Hídrico”, no plano de ação de “Organização de dados e informações para suporte a programas de Pagamento por Serviços Ambientais hídricos”. A partir dessa base de dados, embora pequena, foram realizados testes para textos em inglês e português separadamente. A vantagem da base de dados ser pequena é que os resultados podem ser subjetivamente avaliados.

A primeira ferramenta utilizada neste trabalho para análise textual foi a TaxEdit (MOURA et al., 2012). Para que os objetivos do processo sejam cumpridos, ou seja, a extração e organização de conhecimento, a metodologia de mineração de textos ocorre em um conjunto de passos. A primeira etapa consiste em estruturar os textos em um formato adequado para extração de conhecimento e é chamada de pré-processamento. Nesta etapa os textos são convertidos em uma forma plana e sem formatação. Além disso, ocorre a seleção de termos, ou seja, a extração de um conjunto de termos significativos. Neste passo as stopwords (palavras que, estatisticamente, podem ser consideradas como não influentes na análise da coleção de textos) são removidas e as restantes tem seus sufixos removidos (processo de stemming), dessa forma, casos de multiplicidade de palavras (com a mesma forma pós remoção dos sufixos) são tratados como um único termo. A seleção dos termos pode ser feita estatisticamente, por meio da distribuição de suas frequências no conjunto de textos. A seguir, o software trabalha para organizar os textos de acordo com a proximidade de seus termos, por exemplo, e procura apresentar essa organização por meio de métodos de agrupamento, como o agrupamento hierárquico aglomerativo. Essas etapas podem ser realizadas várias vezes, de acordo com o interesse de quem executa o processo. A cada iteração a lista de stopwords pode ser aprimorada e, consequentemente, o agrupamento é alterado. Finalmente, pode-se usar ferramentas gráficas para apresentar os resultados obtidos.

Na Figura 1 são mostrados alguns resultados obtidos no processo de mineração de textos sobre pagamentos por serviços ambientais hídricos. Na Figura 1, ao fundo, podem-se observar alguns tópicos encontrados na coleção de artigos em inglês pela ferramenta TaxEdit. Ao se clicar em um nó, pode-se observar a frequência de uma palavra-chave do tópico no grupo de documentos indicados. A frequência para a coleção toda pode ser observada no gráfico apresentado em destaque na Figura 1. Nesse caso, praticamente todos os artigos que contém o termo "carbono" estão sob o tópico (grupo) indicado, que por sua vez trata-se de um sub-tópico de controle de diversidade, que está sob oportunidades de negócios, sob modelos de práticas protecionistas – conforme hierarquia de tópicos mostrada.

Futuramente, será realizada uma série de análises textuais com a mesma base de dados (textos de pagamentos por serviços ambientais hídricos) utilizando, porém, os sistemas de software Torch (MARCACINI; REZENDE, 2010) e Mallet (MCCALLUM, 2002). Espera-se, inclusive, poder avaliar o de-

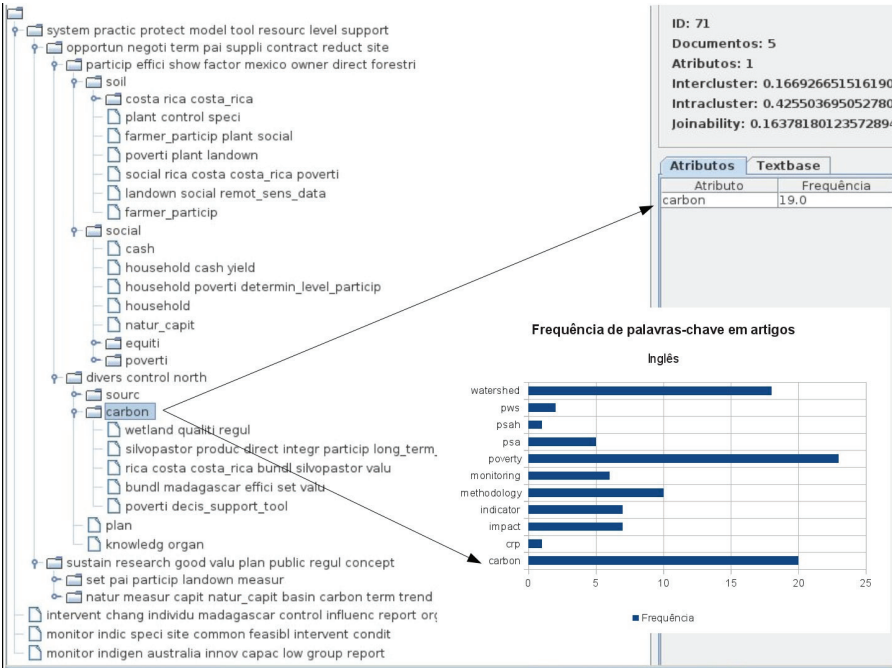

Figura 1. Exemplo de hierarquia de tópicos e gráfico de frequência.

sempenho das ferramentas através de comparações dos resultados obtidos nas análises. E, principalmente, poder observar as tendências dos tópicos encontrados no tempo e em regiões geográficas específicas, para auxiliar processos de tomada de decisão em priorização de áreas para estabelecimento de programas de pagamento por serviços ambientais.

Referências

MARCACINI, R. M.; REZENDE, S. O. Torch: a tool for building topic hierarchies from growing text collection. In: WORKSHOP ON TOOLS AND APPLICATIONS, 9.; BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, 8., 2010, Belo Horizonte. **Proceedings**... Belo Horizonte: UFMG, 2010. p. 133-135. Webmedia.

MCCALLUM, A. K. "**Mallet**: a Machine Learning for Language Toolkit." 2002. Disponível em: <http://mallet.cs.umass.edu>. Acesso em: 27 set. 2013. 2002.

MOURA, M. F.; MERCANTI, E.; PEIXOTO, B. M.; MARCACINI, R. M.; TAMADA, T.; LIMA, A. F.; SANTOS, F. F. dos; REZENDE, S. O.; HIGA, R. H. **TaxEdit - Taxonomy Editor V 3.0**. Versão 1.0. Campinas: Embrapa Informática Agropecuária, 2012. 1 CD-ROM.