Análise histórica de tópicos de publicações em agroinformática

Bruno Malveira Peixoto[1]
Maria Fernanda Moura[2]

Uma forma de analisar o progresso científico em uma área específica do conhecimento ao longo dos anos é observar os tópicos abordados em diversos congressos. Encontrar semelhanças entre artigos publicados, no entanto, exige um trabalho de agrupamento de artigos e identificação dos tópicos abordados. Algoritmos de *clusterização* auxiliam este processo, facilitando a identificação de tópicos comuns entre os artigos analisados.

Assim, neste trabalho supõe-se que é possível analisar a produção científica mundial em tecnologia da informação para o domínio agropecuário e compará-la com as pesquisas realizadas na Embrapa Informática Agropecuária, considerando os anais do congressos *European Federation for Information Technology in Agriculture* (EFITA), *Asian Federation for Information Technology in Agriculture* (AFITA), *World Congress on Computers in Agriculture* (WCCA) e o Congresso Brasileiro de Agroinformática (SBIAgro).

Para realizar esse tipo de análise de dados foi utilizado o processo de mineração de textos proposto por Moura et al. (2008). Esse processo visa à identificação do vocabulário do domínio presente na coleção

[1] *Universidade Estadual de Campinas, UNICAMP; brunomp@cnptia.embrapa.br*

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

de textos, utiliza um algoritmo de agrupamento hierárquico e gera automaticamente uma taxonomia de tópicos, que pode ser editada pelo usuário final com o auxílio de medidas estatísticas de validação das edições realizadas.

Na implementação prática desse processo foram utilizadas duas ferramentas, PreText II (SOARES et al., 2008) e Taxtools (MORETTI et al., 2010). A estimativa inicial do vocabulário de domínio é realizada para o pré-processamento dos dados com a ferramenta PreText II, o vocabulário é assumido como as palavras identificadas pela ferramenta, que podem ser simples ou compostas e são denominadas de n-gramas. Ainda, essas palavras podem ser *stemizadas* ou não, isto é, pode-se remover suas inflexões. Após a identificação dessas palavras, podem ser aplicados filtros a partir de gráficos de frequência, eliminando-se as muito comuns ou muito raras na coleção de artigos. Com isso, obtém-se a padronização dos textos, criação das tabelas atributo-valor referentes a cada coleção de artigos, que foram separados de acordo com o congresso e o ano de publicação. A seguir, utiliza-se a ferramenta Taxtools para obter uma *clusterização* hierárquica dos documentos e realizar cortes (MARCACINI et al., 2009) nos agrupamentos obtidos, para facilitar a visualização dos resultados. Esse tipo de corte de agrupamento é estudado por Marcacini et al. (2009), para o qual é realizada uma junção de grupos de baixo para cima na árvore de documentos, ou seja, uma espécie de *"pruning"* da árvore. Após esses cortes, procura-se identificar descritores específicos de cada grupo com o algoritmo *Robust Labeling Up Method* (RLUM) (MOURA et al., 2010). A partir desses descritores, é possível identificar categorias e/ou assuntos específicos da coleção, que podem ser re-especificados na etapa de pré-processamento e obter resultados melhorados. O algoritmo RLUM foi incorporado à Taxtools no escopo deste trabalho (MORETTI et al., 2010).

Um resultado parcial de uma análise é mostrado na Figura 1. Foram utilizados 108 artigos em inglês do congresso EFITA do ano 2001, obtidos do site do evento. A etapa de pré-processamento selecionou palavras que, em toda a coleção, apareciam com uma frequência entre 17 e 186 e em no mínimo 2 artigos diferentes. Considerou-se no

mínimo 2 artigos para que essas palavras possam ser significativas na formação de um grupo (*cluster*), com mais de um artigo.

A seguir, na Taxtools, foram calculadas medidas de intercluster e fusão para auxiliar no processo de corte de agrupamentos, que foram realizados com valores 0.2 e 0.3 respectivamente, como sugerido por Marcacini et al. (2009). Após os cortes, foi utilizado o algoritmo RLUM.

Na Figura 1.a é mostrado um agrupamento de artigos sobre imagens de mapas. Percebe-se que a palavra *'map'* aparece em todos os nós do ramo da árvore. Após a identificação de descritores, na Figura 1.b *'map'* aparece apenas no nó mais representativo desse agrupamento, retirando informações redundantes. Cada artigo dentro do agrupamento é representado por um conjunto de suas palavras específicas mais frequentes.

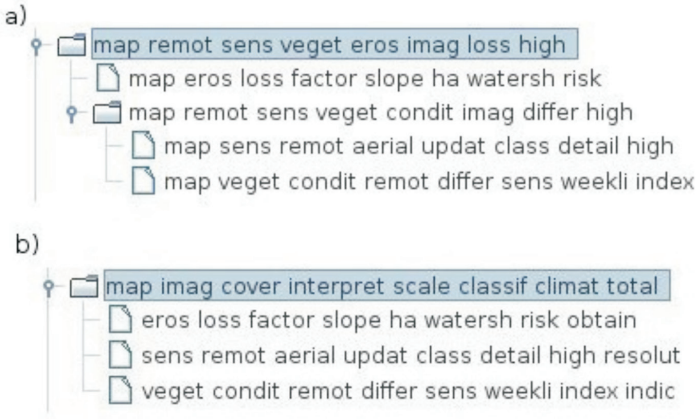

Figura 1. Nós de uma hierarquia: a) Resultado inicial do agrupamento b) Nós equivalentes após os processos de corte e rotulação.

Futuramente, será realizada uma análise de tópicos identificados nas hierarquias geradas, comparando os diversos congressos entre si e com as publicações da Embrapa, apontando temas comuns e divergentes, bem como a evolução dos tópicos ao longo dos anos. Para isso, prevê-se acrescentar a cada nó da árvore o gráfico descritivo de evolução do tópico ao longo dos anos.

Referências

MARCACINI, R. M.; MOURA, M. F.; REZENDE, S. O. Uma abordagem para seleção de grupos significativos em agrupamento hierárquico de documentos. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL, 2009, Bento Gonçalves. **Anais**... Porto Alegre: UFRGS, 2009. p. 1-16. 1 CD-ROM.

MORETTI, C. J. P.; PEIXOTO, B. M.; MOURA, M. F. Tutorial da TaxTools. Campinas, 2010. A ser editado pela Embrapa Informática Agropecuária. (Série. Documentos).

MOURA, M. F.; MARCACINI, R. M.; NOGUEIRA, B. M.; CONRADO, M. S.; REZENDE, S. O. A proposal for building domain topic taxonomies. In: WORKSHOP ON WEB AND TEXT INTELLIGENCE, 1.; SIMPÓSIO BRASILEIRO DE INTELIGÊNCIA ARTIFICIAL, 19., 2008, Salvador, **Proceedings...** São Carlos, SP : ICMC/USP, 2008. v. 1, p. 83-84.

MOURA, M. F.; REZENDE, S. O. A Simple Method for Labeling Hierarchical Document Clusters. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND APPLICATIONS, 10., 2010, Innsbruk - Austria. **Proceedings...** Acta Press, 2010. p. 336-371. v. 1.

SOARES, M. V. B.; PRATI, R. C.; MONARD, M. C. PreText II: descrição da reestruturação da ferramenta de pré-processamento de textos. São Carlos, SP: USP, ICMC, 2008. 45 p. (Relatório técnico, n. 333).