Extração de palavras-chave e taxonomias de tópicos para a BDPA

Claudia Juliana Poker Moretti[1]
Maria Fernanda Moura[2]

Este trabalho refere-se ao desenvolvimento de técnicas e ferramentas baseadas em extração de informação e mineração de textos para análise automatizada de conteúdo de campos dos registros existentes na Base de Dados de Pesquisa Agropecuária (BDPA). A princípio, os objetivos são chegar a uma taxonomia de tópicos inicial para sub-bases da Produção Científica da Embrapa (ProdEMB) e o estudo e identificação de campos faltantes nos registros da BDPA.

Para atingir o primeiro objetivo vem sendo utilizada a metodologia TopTax (MOURA et al. 2008a, 2008b) com seu arcabouço de ferramentas automáticas. A metodologia foi aplicada às sub-bases da ProdEMB, com o objetivo de identificar tópicos e sub-tópicos desses domínios. Nessa tarefa identificam-se candidatos a termos de domínio e vocabulário controlado, o que permite que se avalie a possibilidade de expandir os *thesaurus* utilizados ou a criação da versão inicial de algum deles.

A mineração de textos é composta por cinco etapas: 1. identificação do problema, onde é decidido o que será estudado e porquê; 2. pré-processamento, em que, é feita a escolha de atributos e a

[1] *Universidade Estadual de Campinas; claudiajpm@cnptia.embrapa.br*

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

transformação de dados textuais em dados numéricos; 3. extração de padrões, que é a etapa em que algum modelo de aprendizado de máquina é aplicado aos textos; 4. pós-processamento e 5. utilização do conhecimento.

Neste trabalho, para a etapa de pré-processamento, vem sendo utilizada a ferramenta PreText II (SOARES et al., 2008) e, para a etapa de extração de padrões, a ferramenta TaxTools (MORETTI et al., 2010). A PreText II transforma uma coleção de documentos em uma representação matricial, tendo nas linhas os documentos, nas colunas os atributos e em cada célula um valor de associação do atributo ao documento (pode ser frequência, *tf-idf*, etc.). Os atributos são palavras simples ou compostas, ditas n-gramas, que podem ser *stemmizadas*. A PreText permite que se executem filtros de atributos tais como: a retirada de *stopwords*, frequência mínima e máxima de cada atributo na coleção e/ou em documentos da coleção. O processo de *stemmização* utilizado tem base no algoritmo de Porter (1980) e foi adaptado para as línguas Português, Inglês e Espanhol. A TaxTools é uma ferramenta que permite que se gere, a partir da representação matricial obtida pela PreText: agrupamentos hierárquicos de documentos (*bottom up*, utilizando similaridade de cosseno e o algoritmo "*average*"); cortes do agrupamento, com base em fusão e variância inter e intra-grupo (MARCACINI et al., 2009); descritores estatisticamente mais significantes para cada grupo (MOURA; REZENDE, 2010); e, visualização dos resultados por uma *foldertree* associada a hiper-textos. A ferramenta foi desenvolvida no Instituto de Ciências Matemáticas e de Computação (ICMC), da Universidade de São Paulo (USP) e tem sido mantida e evoluída na Embrapa Informática Agropecuária, da Empresa Brasileira de Pesquisa Agropecuária (Embrapa).

Para o estudo das ferramentas e adaptações das mesmas, foi utilizada uma sub-base da ProdEMB. Na Figura 1, é mostrado o resultado final da TaxTools após a fase de pré-processamento e agrupamento. Foi utilizada a PreText II com a medida *tf* e os filtros de atributos por frequência para gerar as matrizes que serviram de entrada para a TaxTools. Utilizando a ferramenta TaxTools, foi realizada a *clusteri-*

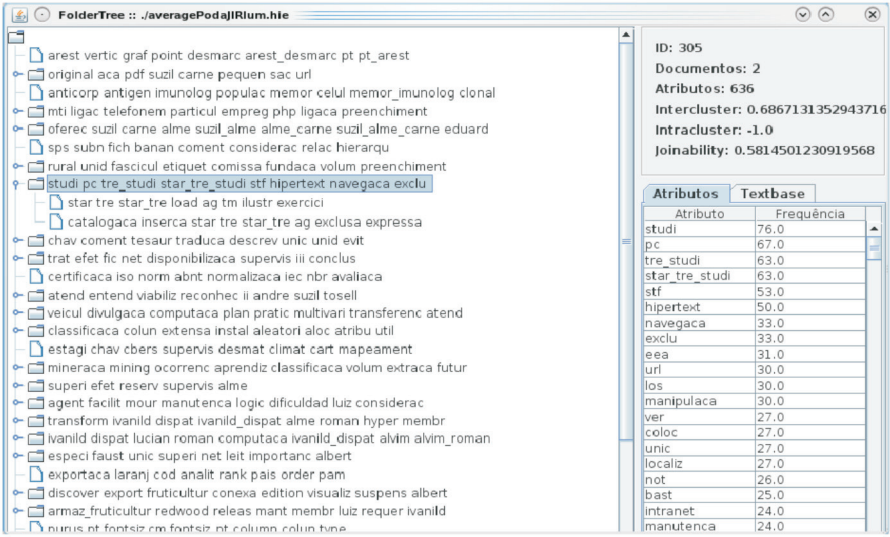

Figura 1. Visualização da etapa de extração de padrões (sub-base Pro-dEMB)

zação da matriz de atributos, o cálculo das medidas, os cortes dos agrupamentos e foi aplicado o *Robust Labelling Up Method* - RLUM (MOURA; REZENDE, 2010) para a identificação de assuntos e/ou categorias. Ainda, no resultado apresentado na Figura 1, observa-se um nó grifado, sobre navegação em hipertextos. Deve-se notar que o tópico mais genérico é “navegação e hipertextos”, sob o mesmo têm-se os assuntos relacionados, “*star-tree*”, “catalogação”, etc.

Esses resultados serão utilizados para atribuir ou sugerir termos para indexação de registros existentes na BDPA, bem como de novos registros. Dessa forma, pode-se complementar ou preencher campos faltantes nos registros. Ainda, com uma taxonomia de tópicos validada, pelas bibliotecárias responsáveis pela BDPA, pode-se gerar classificadores estáticos e/ou dinâmicos para novos documentos.

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 : USP, ICMC, 2008, p. 83-84. 2008-a. v. 1.

MOURA, M. F.; MARCACINI, R. M.; NOGUEIRA, B. M.; CONRADO, M. S.; REZENDE, S. O. **Uma abordagem completa para a construção de taxonomias de tópicos em um** d**omínio**. São Carlos, SP: USP, ICMC, 2008b. (Relatório técnico, n. 329).

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.

PORTER, M.F. An algorithm for suffix stripping, **Program**, v. 14, n. 3, p. 130−137, 1980.

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).