COMPARAÇÃO DE FERRAMENTAS DE TÓPICOS EM TEXTOS SOB O PARADIGMA DE APRENDIZADO ESTATÍSTICO NÃO SUPERVISIONADO

Ivan Prado da **Costa**[1]; Maria Fernanda **Moura**[2]

Nº 14602

RESUMO *- O processo de Mineração de Textos, um caso particular de Mineração de Dados, é composto por cinco etapas: identificação do problema e objetivos; pré-processamento; extração de padrões; pós-processamento e aplicação dos resultados (uso do conhecimento).*

Neste trabalho pretende-se estudar e explorar ferramentas de aprendizado de tópicos hierárquicos, através de coleções textuais e teoria estatística. A modelagem de tópicos nos fornece um conjunto de algoritmos para descobrir estruturas de coleções textuais, ou seja, induzir grupos a partir dos dados observados. Os resultados desses algoritmos podem ser usados para resumir, visualizar e explorar os dados. Para realizar esses procedimentos existem algumas ferramentas de domínio público que têm sido bem avaliadas por grupos de Pesquisa, Desenvolvimento e Inovação. Inicialmente foram utilizadas as ferramentas TaxEdit e TORCH, variando as formas de pré-processamento e obtenção dos tópicos. O experimento inicial baseia-se no trabalho com bases de dados distintas para que o processo possa ser estudado através de comparações. A primeira coleção de textos está em inglês e trata de PSA (Pagamento por Serviços Ambientais) Hídricos, enquanto que a segunda base é um conjunto de notícias agrícolas em português coletadas entre os meses de fevereiro e março de 2014. De acordo com o domínio dos textos, é possível utilizar vocabulário controlado, isto é, uma coleção de termos mantida e atualizada por alguém com conhecimento do assunto. Os tópicos encontrados foram bem avaliados e utilizados como suporte para outros trabalhos. Além disso, nota-se diferença entre experimentos realizados com e sem o uso do vocabulário controlado.

Palavras-chaves: Mineração de Textos, tópicos hierárquicos, comparação, análise estatística.

1 Autor, Bolsista CNPq (PIBIC): Graduação em Estatística, UNICAMP, Campinas-SP; ivanpradoc@gmail.com
2 Orientador: Pesquisador da Embrapa Informática Agropecuária, Campinas-SP; maria-fernanda.moura@embrapa.br.

ABSTRACT- *The process of Text Mining, a particular case of Data Mining, consists of five steps: identification of problem and objectives; pre-processing; extraction of patterns; post-processing and application of results (use of knowledge).*

This paper aims to study and explore tools of learning hierarchical topics through collections of texts and statistical theory. The modeling of topics provides us a set of algorithms for discovering structures of textual collections, namely, inducing groups from the observed data. The results of these algorithms can be used to summarize, visualize and explore the data. To perform these procedures, there are some tools in the public domain that have been well evaluated by groups of Research, Development and Innovation. Initially it is used the TaxEdit and TORCH tools with various forms of pre-processing and obtaining of topics. The initial experiment is based on work with different databases so that the process can be studied by comparison. The first collection of texts is in English and comes to PSA (Payment for Ecosystem Services) Hydric, while the second base is a set of agricultural news in Portuguese collected between the months of February and March 2014. According to the domain of the texts, it is possible to use controlled vocabulary, namely a collection of terms maintained and updated by someone with knowledge of the matter. The topics found were well evaluated and used as a support for other jobs. Furthermore, there is difference between experiments with and without the use of controlled vocabulary.

Key-words: Text Mining, hierarchical topics, comparison, statistical analysis.