Classificação Automatizada de Textos Científicos em Áreas de Conhecimento do Ensino Superior

Thales V. Maciel[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia Sul-rio-grandense (IFSul)
Campus Bagé – Av. Leonel de Moura Brizola, 2501 – 96408400 – Bagé – RS – Brasil

thalesmaciel@ifsul.edu.br

Abstract. *This paper documents a methodology based in text mining for the automated classification of scientific papers upon a specific taxonomy of knowledge areas. It is proposed that keywords, as part of texts' metadata, were used for algorithm training and respective abstracts used for testing. Test results showed good potential for further application in the real world.*

1. Introdução

No meio acadêmico, é comum a ocorrência de eventos científicos e periódicos multidisciplinares. Neles, trabalhos de diversas áreas de conhecimento devem passar pela avaliação do respectivo mérito científico antes de serem aprovados para publicação nos anais do evento ou periódico.

Neste contexto, comumente são designados agentes humanos como avaliadores para cada trabalho submetido. Para prover eficiência na avaliação, esta designação ocorre através da associação manual entre as áreas de conhecimento informadas por cada avaliador como de seu interesse e aquela informada pelos autores dos trabalhos a qual a obra esteja supostamente inserida.

Ocorre que, muitas vezes, a informação da área de conhecimento dos trabalhos submetidos é incompleta, incorreta ou ausente, o que causa transtornos quando da avaliação pelos avaliadores originalmente designados ou, até mesmo, da própria designação de avaliadores para tais submissões.

O presente trabalho teve a problemática definida na forma de "como automatizar a atribuição de áreas de conhecimento para textos científicos?". A hipótese estudada é a de que as palavras-chave cadastradas como metadados dos textos podem ser utilizadas nesta automatização, especificamente em tarefas de mineração de texto. Assim, é objetivada a caracterização de um método constante de viabilidade prática no que se refere ao consumo de recursos computacionais, tempo de execução e âmbito aceitável para a ocorrência de erros de classificação para predição da área de conhecimento referente a textos científicos.

2. Solução Proposta

Frank, Hall e Witten (2016) definem a mineração de texto como o processo de análise de texto automatizada que visa a extração de informação útil para fins específicos. Estes autores classificam a mineração de texto como uma especialização da mineração de dados, diferenciando-as em que, no caso de texto, a informação é explícita, embora não estruturada.

O conjunto de dados disponibilizado para esta análise foi composto por 997 instâncias de trabalhos científicos, todas constituídas por dois atributos: um representando o conjunto de quatro palavras-chave distintas, que são comumente requisitos para submissão de trabalhos científicos e utilizadas para fins de indexação de obras, e outro representando a respectiva classificação na árvore de áreas de conhecimento proposta pela CAPES (2008) em seu nível mais específico.

Para fins da realização dos experimentos com mineração de texto sobre o conjunto de dados descrito, foi utilizado o Waikato Environment for Knowledge Analysis (WEKA), que é uma coleção de algoritmos que podem ser utilizados em atividades de mineração de dados diversas, como classificação, regressão, associação,

clustering e mineração de texto, além de pré-processamento de conjunto de dados e visualização de resultados (Hall et al. 2009).

A experimentação central, descrita neste estudo, foi desempenhada com a utilização de classificação filtrada, que possibilita a aplicação de técnicas de filtragem ou seleção de dados em conjunto com a execução de um algoritmo de classificação ou regressão, estendendo o aprendizado realizado por tal algoritmo ao produto do melhoramento na qualidade dos dados, porém, sem alterar o estado do conjunto de dados original (Frank, Hall e Witten, 2016).

Outrossim, a atividade de classificação filtrada foi parametrizada para realizar a conversão do atributo referente ao conjunto de palavras-chave dos trabalhos em um novo conjunto de atributos, então representando a ocorrência ou não de cada palavra-chave contida no conjunto de dados original em cada instância de trabalho. Esta atividade de filtragem, por sua vez, foi configurada, a partir de seu comportamento padrão, para formatar as palavras-chave em letras minúsculas exclusivamente e ignorar a ocorrência de palavras contidas em uma lista auxiliar, onde estavam contidas expressões julgadas irrelevantes ao contexto de classificação de textos científicos.

O algoritmo efetivamente responsável pelo processamento na mineração de dados foi o Naive Bayes Multinominal (Mccallum e Nigam, 1998), projetado com vistas no aprendizado de máquina para classificação de texto. Este algoritmo é implementado no WEKA sob nome MultinomialNaiveBayes (Frank, Hall e Witten, 2016), onde não há parametrização ao seu comportamento padrão.

A etapa de testes do modelo descoberto foi realizada com um conjunto de dados correspondente aos textos dos resumos dos mesmos trabalhos cujas palavras-chave foram utilizadas no treinamento do algoritmo. O método obteve 69,007% de acurácia nas classificações dos trabalhos e o valor de 0,67 para o coeficiente de kappa. Valores do coeficiente de kappa neste âmbito indicam evidências substanciais de que os resultados obtidos não foram apenas casuais, apresentando, de fato, relevância estatística para o domínio de aplicação (Vieira e Joanne, 2005).

3. Considerações Finais

Este trabalho possibilitou a automatização da atribuição das áreas de conhecimento de trabalhos científicos com base em uma aplicação do aprendizado de máquina, na forma de mineração de texto. Outrossim, foi julgado como satisfeito, o objetivo desta pesquisa.

Trabalhos futuros envolvem o tratamento do desbalanceamento entre as classes do conjunto de dados original e a investigação de um método para obter proveito da estrutura de árvore sob a qual estão organizadas as áreas de conhecimento no Brasil.

Referências

Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (2008) “Portaria nº 09/2008”, http://reality.sgi.com/employees/jam_sb/mocap/MoCapWP_v2.0.html, December.

Frank E., Hall, M. A., and Witten, I. H. (2016). The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.

Hall M., Frank E., Holmes G., Pfahringer, B., Reutemann P. and Witten I. (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, Volume 11, I 1.

Mccallum A. and Nigam K. (1998) A Comparison of Event Models for Naive Bayes Text Classification. In: AAAI-98 Workshop on 'Learning for Text Categorization'.

Viera, A. and Joanne M. "Understanding interobserver agreement: the kappa statistic." Fam Med 37.5 (2005): 360-363.