Classificação e visualização de dados de patentes

Claudia A. Martins, Rafaela S. Francisco, Henrique C. Farias

Instituto de Computação – Universidade Federal de Mato Grosso (UFMT)
Av. Fernando Corrêa da Costa, n. 2367, Boa Esperança, Cuiabá-MT – Brazil

claudia@ic.ufmt.br, {rafaela.souzaf,harrycamachofarias}@hotmail.com

Abstract. *This paper describes the methodology to assist the patent data classification process. In addition to the application of document classification and vectorization algorithms, visualization techniques are being investigated to assist in all stages of the classification process.*

Resumo. *Este artigo descreve uma metodologia para auxiliar o processo de classificação de dados de patentes. Além da aplicação dos algoritmos de classificação e vetorização dos documentos, estão sendo investigadas técnicas de visualização para auxiliar em todas as etapas do processo de classificação.*

1. Introdução

Processamento de dados baseados em textos é uma tarefa complexa que demanda técnicas computacionais tais como Processamento da Linguagem Natural (PLN) e Aprendizado de Máquina (AM) que, entre suas diversas aplicações, podem ser utilizadas no processamento automático para análise, compreensão e manipulação dos dados textuais. Uma das áreas de aplicação de processamento de textos complexa pela natureza intrínseca de seu conteúdo é a classificação automática de documentos de patentes. De acordo com o INPI[1], "uma patente consiste em um título de propriedade temporária sobre uma invenção ou modelo de utilidade, outorgado pelo Estado aos inventores, ou autores, detentoras de direitos sobre a criação".

As patentes são armazenadas de forma hierarquizadas em categorias de acordo com as características de seu conteúdo, como área médica ou alimentícia, e cada escritório define qual sistema de classificação utilizar. O *International Patent Classification* (IPC) (WIPO 2019) é um dos sistemas de classificação, cuja organização se baseia numa hierarquia de níveis e subníveis, utilizado em mais de noventa países e abrange as áreas tecnológicas (Fall et al. 2003).

O processo de classificação e visualização de dados de patentes podem ser usadas como ferramentas de busca e recuperação de conteúdos para auxiliar na prospecção de oportunidades e, também, para garantir que novos depósitos de patentes não infrinjam leis de propriedade intelectual. Dessa forma, o objetivo desse trabalho é análise e automatização de ferramentas que possam auxiliar na aplicação de algoritmos vetorização, classificação e visualização de informações de patentes (Grawe et al,2017; Zhang, 2011; Dietterich, 1998).

1 https://www.gov.br/inp

2. Metodologia

Os escritórios de patentes possuem formas específicas e utilizam diferentes sistemas de classificação para o armazenamento de patentes. O sistema de classificação IPC é um sistema hierarquizado em diferentes níveis ou camadas. Para que uma patente possa ser classificada em um nível, a patente precisa obrigatoriamente ter sido classificada ao nível anterior. A complexidade no processo de classificação de dados de patentes consiste, além da dimensionalidade inerente dos dados e da sobreposição das classes, também a especialização do conteúdo nessa hierarquia de níveis, pois à medida que vai percorrendo na hierarquia, o assunto ou área que a patente abrange, vai se especializando e assumindo, consequentemente, diversos níveis de classificações, conhecido como um problema multirótulo.

De forma geral, a metodologia utilizada consiste no fluxo de processamento ilustrado na Figura 1.

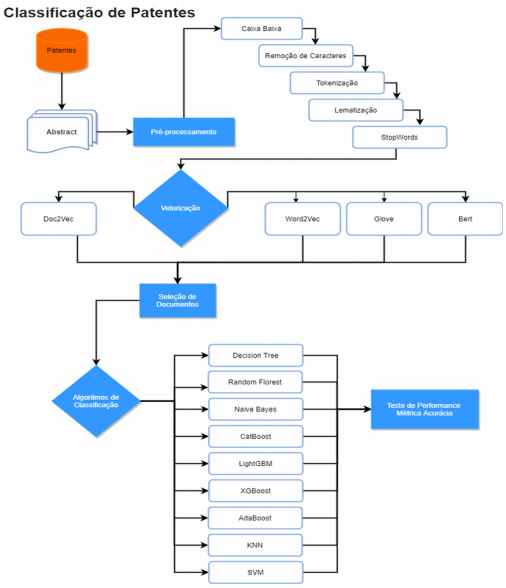

Figura 1: Metodologia

Resumidamente, a partir de um conjunto de dados textuais, os documentos são pré-processados e vetorizados utilizando técnicas de *word embedding*. Após, a representação dos documentos são processados com algoritmos de classificação. Em qualquer uma dessas etapas, técnicas estatísticas e de visualização estão sendo aplicadas tanto para melhorar a compreensão geral dos dados quanto para identificar termos mais relevantes que possam auxiliar no refinamento do processo de classificação das patentes (Whitehead, 2017; Manning, 1999).

Neste trabalho, inicialmente, foram selecionados 75.239 documentos disponíveis pela WIPO-alpha[2], divididos entre treinamento e teste, classificados em 8 seções – de A a H. A quantidade de documentos nas classes A-H está totalmente totalmente desbalanceada, cuja classe C tem 21,6% de todos os documentos (16.244) e a classe D possui apenas 2,3% (1.710). Para minimizar esse problema, foi realizado um pré-processamento com o objetivo de selecionar os documentos mais relevantes (Farias et. al, 2021).

2 https://wipo-analytics.github.

3. Resultados

A utilização de técnicas de visualização de termos e informações são importantes para auxiliar na compreensão dos dados e, consequentemente, no desenvolvimento de mecanismos automáticos que auxilie o usuário em um processo de busca e extração de termos e expressões em dados textuais.

Alguns trabalhos já foram desenvolvidos com o objetivo de analisar o desempenho dos algoritmos de classificação, juntamente com os algoritmos de vetorização de palavras (Farias et. al, 2021). O melhor resultado dos experimentos foi obtido pelo algoritmo de classificação Light GBM com o vetorizador Word2Vec, cuja acurácia foi de 82,74% no conjunto de treino e 83,36% no conjunto de teste. É um resultado similar e até superior aos encontrados na literatura (Jafery, 2019; Lyu, 2019).

Atualmente, a incorporação de novas funcionalidades estatísticas e de visualização estão nesta segunda rodada de experimentos, visando refinar e melhorar a acurácia dos classificadores no sentido de verificar a correspondência, por exemplo, entre as palavras mais frequentes no conjunto de dados e as palavras mais relevantes em cada classe; quais palavras relevantes estão nos limites entre as classes; qual a necessidade de balanceamento de representatividade de palavras e documentos na discriminação das classes.

Dessa forma, técnicas de redução de dimensionalidade, juntamente com técnicas de *clustering* visual de textos, *cloudwords* e TF-IDF estão em fase implementação para encontrar as palavras mais relevantes em cada classe.

4. Conclusão

Muitas vezes, uma representação visual dos dados se torna um mecanismo importante no processo de análise, compreensão e mesmo redução dos dados. Assim, a aplicação de algoritmos que buscam por padrões estatísticos e visuais na identificação de termos, visam auxiliar o processo de classificação das patentes. A incorporação dessas técnicas e ferramentas em novas funcionalidades no processamento de classificação de documentos consistem em melhorar a recuperação de informações considerando as especificidades das possíveis definições de uma patente, o seu contexto e os termos envolvidos. Este trabalho está incorporando essas funcionalidades no processo de classificação de dados de patentes.

Agradecimento

Ao apoio financeiro da Fundação de Amparo à Pesquisa do Estado de Mato Grosso (FAPEMAT) - Projeto n.0213429/2017.

Referências

Dietterich, T. G. (1998) Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10(7):1895–1923, 1998.

Fall, C. J.; Törcsvári, A.; Benzineb, K.; Karetka, G. (2003) “Automated categorization in the international patent classification”, ACM SIGIRForum(37:1), pp. 10–25. URL http://portal.acm.org/citation.cfm?doid=945546.945547

Farias, H. C.; Bonfante, A. G.; Martins, C. A. (2020) Seleção de documentos baseado em centroides para classificação de patentes usando Word2Vec e KNN. Anais do XLVII Seminário Integrado de Software e Hardware (SEMISH), CSBC, 2021. DOI: https://doi.org/10.5753/semish.2021.

Farias, H. C.; Martins, C. A.; Francisco, R. S. (2021) Algoritmos de Classificação e Representação *Word Embedding* em Dados de Patentes. Proceedings 5th Conference on Information Systems in Latin America (ISLA 2021).

Grawe, M. F.; Martins, C. A.; Bonfante, A. G. (20217) Automated Patent Classification Using Word Embedding. 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, Cancun. DOI: 10.1109/ICMLA.2017.0-127.

Jafery, W. A. Z. W. C., Omar, M. S. S., Ahmad, N. A., Ithnin, H. (2019) "Classification of Patents according to Industry 4.0 Pillars using Machine Learning Algorithms", in 2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS), IEEE.

Lyu, L.; Han, T. (2019) "A comparative study of Chinese patent literature automatic classification based on deep learning", in Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, vol. 2019-June, Institute of Electrical and Electronics Engineers Inc., vol. 2019-June.

Manning, C. D; Schutze, H. (1999) Foundations of statistical natural language processing. MIT Press, Cambridge, USA, 1999.

WIPO (2019). Guide to the International Patent Classification. Technical report. https://www.wipo.int/portal/en/index.html.

Whitehead, M.; Johnson, D. K. N. (2017) A Tool for Visualizing and Exploring Relationships among Cancer-Related Patents. Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference 2017, Association for the Advancement of Artificial Intelligence (www.aaai.org), pages 235-238, 2017.

Zhang, Q.; Segall, R. S.; Cao, M. (2011) Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications, IGI Global, Hershey, PA, 2011.