UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO

Henrique Cursino Vieira

Modelo para Trabalho de Conclusão de Curso em LaTeX utilizando a classe USPSC para o ICMC

São Carlos

2020

Henrique Cursino Vieira

Modelo para Trabalho de Conclusão de Curso em LaTeX utilizando a classe USPSC para o ICMC

Trabalho de conclusão de curso apresentado ao Centro de Ciências Matemáticas Aplicadas à Indústria do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, como parte dos requisitos para conclusão do MBA em Ciências de Dados.

Área de concentração: Ciências de Dados

Orientador: Prof. Dr. Francisco Louzada Neto

Versão original

São Carlos

2020

"Nenhuma grande descoberta foi feita jamais sem um palpite ousado."

Isaac Newton

LISTA DE ABREVIATURAS E SIGLAS

SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency

SUMÁRIO

1 INTRODUÇÃO

No desenvolvimento geral de produtos, a pesquisa por documentos de patentes visa garantir o não infringimento de propriedades intelectuais que ainda não estão em domínio público (BREITZMAN; MOGEE, 2002). Em uma pesquisa de documentos de patentes, patentes relacionadas a tecnologia, economia e jurídico são tratadas, classificadas e analisadas para se obter um alto valor técnico e comercial (LI, 2018). De acordo com o *World Intellectual Property Indicator 2017*, em 2016, o numero de documentos de patente excedeu 3 milhões pela primeira vez, um aumento de 8.3% (LI, 2018). Mas como realizar esta tarefa havendo algumas centenas de documentos de patentes sobre um assunto específico? O método tradicional necessita de tempo e equipe para realizá-lo, apresentando um resultado com deficiências devido ao alto volume de documentos de patente a serem analisadas (LI, 2018). Hoje, já há portais web que oferecem ferramentas das quais algumas auxiliam ao pesquisador a reduzir essa pesquisa (ABBAS; ZHANG; KHAN, 2014), mas classificam os documentos em uma relevância geral. Esse resultado somente demonstra que dentro daquela amostra de documentos, uma visão macro sobre o assunto que muitas vezes o pesquisador está em busca de um subassunto, como quais mercados essa tecnologia está presente, quais os processos de produção desta tecnologia ou qual a formulação desse composto.

De acordo com Shahid et al (2019), a classificação de documentos de patentes em assuntos e a atribuição de valor de relevância para estes assuntos, permitindo ao pesquisador filtrar as patentes que o interessa e reduzindo o escopo de analise. Nesse, realizou a construção de uma matrix de valores de term frequency - inverse document frequency (tf-idf), notações e peso ponderado por BM25, que posteriormente foi testado em diferentes classificadores, classificando os documentos de patente em cada assunto.

Seguindo Anne et al (2017), identificou uma matriz de métodos a serem aplicados com os modelos k-Nearest Neighbors (kNN), Support Vector Machine (SVM), Random Forest e J48. Os principais passos para essa pesquisa foram técnicas de seleção de características, com uso de ganho de informação e correlação para efetividade do classificadores.

Destes dois estudos, foi observado que a adição de mais características para os modelos de classificação utilizados, a acurácia foi melhorada (SHAHID et al., 2020). E que obstáculos, como o desbalanceamento dos dados foram atenuados pela adição de novas características (ANNE et al., 2017).

Com o rápido crescimento de documentos de patente, torna-se urgente a questão de automatização da classificação de documentos de patente de forma acurada e rápida (ZHU et al., 2020). Os documentos de patente contem um potencial conhecimento tecnológico

na resolução de problemas no processo de fabricação, nos quais são de grande valor cientifico e tecnológico, no entanto, esse conhecimento está implícito em longos textos (LI, 2018; WANG et al., 2016). A classificação de documentos de patente em categorias ou subcategorias utilizando de modelos de aprendizado de maquina se beneficiaria do uso da extração de características uteis vindas do próprio documento (ANNE et al., 2017). Observa-se que mais de 90% das informações de cientificas e tecnológicas estão em documentos de patente, e sua analise resultaria em decisões de negocio de sucesso (LI, 2018).

O objetivo deste estudo é classificar as patentes em assuntos, subassuntos e determinar suas respectivas relevâncias. Faremos o uso do modelo de classificação baseado em florestas aleatórias, a vantagem desse modelo é o uso como regressão e classificação, além da sua facilidade de interpretação do resultado obtido. Não foi encontrado artigos ou materiais que fizessem essa aplicação para patentes relacionadas ao setor agrônomo, para gerenciamento de patentes, desenvolvimento de produtos e descoberta de mercados. Visto tudo isso, buscamos então determinar qual a acurácia na categorização de patentes com o modelo de classificação baseado em florestas aleatórias aplicado a agronomia.

REFERÊNCIAS

ABBAS, A.; ZHANG, L.; KHAN, S. U. A literature review on the state-of-the-art in patent analysis. **World Patent Information**, Elsevier Ltd, v. 37, p. 3–13, 2014. ISSN 01722190. Disponível em: <http://dx.doi.org/10.1016/j.wpi.2013.12.006>.

ANNE, C. et al. Multiclass patent document classification. **Artificial Intelligence Research**, v. 7, n. 1, p. 1, 2017. ISSN 1927-6974.

BREITZMAN, A. F.; MOGEE, M. E. The many applications of patent analysis. **Journal of Information Science**, v. 28, n. 3, p. 187–205, 2002. ISSN 01655515.

LI, G. A Literature Review on Patent Texts Analysis Techniques. **International Journal of Knowledge and Language Processing**, v. 9, n. 3, p. 1–15, 2018.

SHAHID, M. et al. Automatic patents classification using supervised machine learning. In: SPRINGER. **International Conference on Soft Computing and Data Mining**. [S.l.], 2020. p. 297–307.

WANG, G. et al. Extraction of Principle Knowledge from Process Patents for Manufacturing Process Innovation. **Procedia CIRP**, The Author(s), v. 56, p. 193–198, 2016. ISSN 22128271. Disponível em: <http://dx.doi.org/10.1016/j.procir.2016.10.053>.

ZHU, H. et al. Patent automatic classification based on symmetric hierarchical convolution neural network. **Symmetry**, v. 12, n. 2, p. 1–12, 2020. ISSN 20738994.