

**UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE CIÊNCIAS MATEMÁTICAS E DE COMPUTAÇÃO**

Henrique Cursino Vieira

Classificação de patentes utilizando random forest

São Carlos

2020

Henrique Cursino Vieira

Classificação de patentes utilizando random forest

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. Nikolai Valtchev Kolev

Versão original

São Carlos

2020

*“Nenhuma grande descoberta foi feita jamais sem um
palpite ousado.”
Isaac Newton*

LISTA DE ABREVIATURAS E SIGLAS

IDE	Integrated Development Environment
SVM	Support Vector Machine
TF-IDF	Term Frequency - Inverse Document Frequency

SUMÁRIO

1	INTRODUÇÃO	9
1.1	Apresentação	9
1.2	Justificativa	10
1.3	Problema	10
1.4	Objetivo geral	10
1.5	Objetivos específicos	10
1.6	Metodologia	11
	REFERÊNCIAS	13

1 INTRODUÇÃO

1.1 Apresentação

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). O sistema de patentes é um conjunto de medidas utilizados para visar o retorno do valor privado investido ao valor social de suas invenções, fornece aos inventores um período temporário de poder de mercado, recuperando os custos de seus investimentos na pesquisa (WILLIAMS, 2017). De acordo com o *World Intellectual Property Indicator 2017*, em 2016, o número de documentos de patente excedeu 3 milhões pela primeira vez, um aumento de 8.3% (LI, 2018). Em uma pesquisa de documentos de patente, documentos relacionados a tecnologia, economia e jurídico são tratadas, classificadas e analisadas para se obter um alto vantagem técnica e comercial (LI, 2018).

A classificação de documentos é o processo de classificação de um documento em uma categoria predefinida, desempenhando um papel importante no gerenciamento e busca de temas (ANNE et al., 2017). A automatização da classificação de documentos a partir de aprendizado de máquina, pode rotular documentos de um tema único e a rotulagem em vários temas é relativamente desafiador (ANNE et al., 2017).

De acordo com Shahid et al (2019), a classificação de documentos de patente em temas e a atribuição de valor de relevância para estes temas, permitem ao pesquisador filtrar as patentes que o interessa e reduzindo o escopo de análise. Nesse trabalho, realizou a construção de uma matriz 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. Vide 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). Balancear a relação entre esses dois pontos é um desafio quanto a classificação de documentos de patente.

1.2 Justificativa

Como tratar, classificar e analisar documentos de patente havendo algumas centenas de documentos 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 tema que muitas vezes o pesquisador está em busca de um subtema, como quais mercados essa tecnologia está presente, quais os processos de produção desta tecnologia ou qual a formulação desse composto.

1.3 Problema

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 científico e tecnológico, no entanto, esse conhecimento está implícito em longos textos (LI, 2018; WANG et al., 2016). A classificação de documentos de patente em temas e subtemas utilizando de modelos de aprendizado de máquina se beneficiaria do uso da extração de características úteis vindas do próprio documento (ANNE et al., 2017). Observa-se que mais de 90% das informações de científicas e tecnológicas estão em documentos de patente, e sua análise resultaria em decisões de negócio de sucesso (LI, 2018).

1.4 Objetivo geral

Este projeto se propõe a classificar documentos de patente por tema específico e subtemas de interesse do pesquisador, reduzindo o escopo de documentos de patentes a serem estudados à somente os mais relevantes para o que se procura.

1.5 Objetivos específicos

A classificação de documentos de patente envolverá o uso técnicas de processamento de linguagem natural para o tratamento e preparação dos dados que serão usados no modelo de classificação por relevância que será desenvolvido. Este modelo usará inicialmente a medida estatística TF-IDF e avaliaremos outras medidas. Haverá a necessidade de criação de dicionários que auxiliem na classificação dos documentos de patente.

1.6 Metodologia

Realizaremos a obtenção de uma massa de documentos de patente com conteúdo aplicado a agricultura através da ferramenta Google Patents. 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. Faremos o uso do modelo de classificação baseado em florestas aleatórias, a vantagem desse modelo, é a flexibilidade para o uso em regressão e classificação, além da sua facilidade de interpretação do resultado obtido. A construção de dicionários será a partir de técnicas de Processamento de Linguagem Natural, elencando as palavras mais relacionadas a área. A análise, construção de dicionários e modelagem do modelos de regressão e classificação será feita na linguagem de programação R com a integrated development environment (IDE) de desenvolvimento RStudio.

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>>.
- WILLIAMS, H. L. How Do Patents Affect Research Investments? **Annual Review of Economics**, v. 9, n. 1, p. 441–469, 2017. ISSN 1941-1383.
- 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.