



UNIVERSIDADE FEDERAL DE SANTA CATARINA  
CENTRO DE CIÊNCIAS, TECNOLOGIAS E SAÚDE DO CAMPUS ARARANGUÁ  
CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

Davide Clode da Silva

**UM MÉTODO VOLTADO À RECOMENDAÇÃO DE EXAMINADORES  
PARA A ANÁLISE DE PATENTES**

Araranguá

2022

Davide Clode da Silva

**UM MÉTODO VOLTADO À RECOMENDAÇÃO DE EXAMINADORES  
PARA A ANÁLISE DE PATENTES**

Trabalho de Conclusão do Curso de Graduação em Engenharia de Computação do Centro de Ciências, Tecnologias e Saúde do Campus Araranguá da Universidade Federal de Santa Catarina para a obtenção do título de Bacharel em Engenharia de Computação.

**Orientador:** Prof. Alexandre Leopoldo Gonçalves, Dr.

Araranguá

2022

Ficha de identificação da obra elaborada pelo autor,  
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Clode da Silva, Davide

Um Método Voltado à Recomendação de Examinadores para a  
Análise de Patentes / Davide Clode da Silva ; orientador,  
Alexandre Leopoldo Gonçalves, 2022.

35 p.

Trabalho de Conclusão de Curso (graduação) -  
Universidade Federal de Santa Catarina, Campus Araranguá,  
Graduação em Engenharia de Computação, Araranguá, 2022.

Inclui referências.

1. Engenharia de Computação. 2. Análise de Patentes.. 3.  
Sistemas de Recomendação.. 4. Embeddings. 5. Doc2Vec. I.  
Leopoldo Gonçalves, Alexandre. II. Universidade Federal de  
Santa Catarina. Graduação em Engenharia de Computação. III.  
Título.

Davide Clode da Silva

**UM MÉTODO VOLTADO À RECOMENDAÇÃO DE EXAMINADORES  
PARA A ANÁLISE DE PATENTES**

Este Trabalho de Conclusão de Curso foi julgado adequado para obtenção do Título de Bacharel em Engenharia de Computação, e foi aprovado em sua forma final pelo Curso de Engenharia de Computação.

Araranguá, 17 de Março de 2022.

---

**Prof<sup>a</sup>. Analúcia Schiaffino Morales, Dr<sup>a</sup>.**  
Coordenadora do Curso

**Banca Examinadora:**

---

**Prof. Alexandre Leopoldo Gonçalves, Dr.**  
Orientador  
Universidade Federal de Santa Catarina

---

**Prof<sup>a</sup>. Andréa Sabedra Bordin, Dr<sup>a</sup>.**  
Avaliadora  
Universidade Federal de Santa Catarina

---

**Prof. Alessandro Botelho Bovo, Dr.**  
Avaliador  
Universidade Tecnológica Federal do Paraná

# Um Método Voltado à Recomendação de Examinadores para a Análise de Patentes

## *Examiners Recommendation Method For Patent Analysis*

Davide Clode da Silva\* Alexandre Leopoldo Gonçalves†

2022, Março

### Resumo

O processo de análise de patentes é uma tarefa que exige muito trabalho para os examinadores de patentes. Analisar uma patente corretamente diminui, por exemplo, o risco de litígios de patentes. Embora a análise de patentes seja indispensável, o contínuo crescimento no volume de patentes torna o processo de avaliação cada vez mais dispendioso. Entre as possíveis ferramentas com potencial para auxiliar neste desafio estão os sistemas de recomendação, que são usados em diferentes áreas podendo ser empregados para sugerir avaliadores no contexto de análise de patentes. Sendo assim, neste trabalho, propõe-se um método que, no contexto da análise de patentes, possibilite, a partir de uma patente de interesse, a recomendação de um *ranking* de avaliadores com base em patentes previamente avaliadas. No desenvolvimento da etapa de recomendação foi adotada a abordagem de recomendação baseada em conteúdo utilizando o conceito de *embeddings* de documentos por meio do *framework Doc2vec*<sup>®</sup>. Levando-se em conta a avaliação, cinco cenários foram estabelecidos de modo que para cada patente de entrada em cada um dos cenários fosse recomendada uma lista ordenada de possíveis examinadores. Os resultados nestes cenários se mostraram, em sua maioria, consistentes, ou seja, foram capazes de sugerir os examinadores corretos. Por fim, conclui-se que o método proposto viabiliza a recomendação de examinadores com potencial para auxiliar tomadores de decisão em ambientes em que estes precisem vincular determinada patente a um examinador específico.

**Palavras-chave:** Análise de Patentes. Sistemas de Recomendação. *Embeddings*. *Doc2vec*.

---

\*davide.silva@grad.ufsc.br

†alexandre.l.goncalves@grad.ufsc.br

# Um Método voltado à Recomendação de Examinadores para a Análise de Patentes

## *Examiners Recommendation Method For Patent Analysis*

Davide Clode da Silva\* Alexandre Leopoldo Gonçalves†

2022, Março

### **Abstract**

Patent analysis process is a labor intensive task for patent examiners. The correct analysis of a patent, for instance, reduces the risk of patent litigation. Despite the fact that patent analysis is indispensable, the continuous growth in the volume of patents makes the analysis process increasingly expensive. Out of possible tools with the potential to help in this challenge are recommendation systems, which are used in different areas and can be used to suggest examiners in the context of patent analysis. Therefore, in this work, we propose a method that, in the context of patent analysis, allows, from a patent of interest, the recommendation of a ranking of examiners based on a previously evaluated patent. In the development of the recommendation stage, the content-based recommendation approach was adopted using the concept of document embeddings through the Doc2vec® framework. Taking into account the evaluation, five scenarios were established so that it was possible to recommend an ordered list of possible examiners for each input patent in each of the scenarios. The results in these scenarios were mostly consistent, that is, they were able to suggest the correct examiners. Finally, it is concluded that the proposed method enables the recommendation of examiners with the potential to help decision makers in environments where they need to link a certain patent to a specific examiner.

**Keywords:** Patent Analysis. Recommender Systems. Embeddings. Doc2vec.

---

\*davide.silva@grad.ufsc.br

†alexandre.l.goncalves@grad.ufsc.br

# 1 INTRODUÇÃO

O surgimento do termo patente ou patenteamento remonta séculos antes de Cristo. Em um dos projetos de pesquisa aplicada realizados para o rei de Siracusa, Hierão II, no qual este temia que sua nova coroa, que deveria ser de ouro puro, tivesse sido misturada com prata, o grego Arquimedes (287-212 ac.) descobriu que a coroa fora adulterada, ou seja, que na sua composição tinha, além de ouro, a prata. Essa descoberta ocorreu após Arquimedes perceber que, ao comparar o ouro e a prata, ambos com o mesmo peso, o ouro desloca um volume menor de água do que a prata, isto porque o ouro é mais denso do que a prata. Após apresentar sua descoberta ao tribunal de Hierão II, um dos conselheiros do rei impressionou-se com o valor comercial que a descoberta poderia gerar, tendo sugerido que Arquimedes registrasse-a (LATTMAN, 1996).

A patente moderna, datada do século XVIII, é um direito concedido pelo Estado ao inventor e que permite que os direitos deste não sejam violados. Esse direito, às vezes denominado Direito de Propriedade Intelectual (DPI), permite que o inventor dificulte a imitação imediata ou exclua terceiros de fazer ou usar uma determinada invenção e, além disso, é visto como um incentivo à inovação e estimulador de atividade inventiva. O inventor, por sua vez, torna pública a descrição e a implementação da invenção, levando em conta as leis de cada país (HALL, 2007). Para Gorshkov (2021), além da patente atribuir ao seu titular a propriedade, a autoria e o direito de uso exclusivo para produzir, usar e vender invenções no mercado, ajuda a criar barreiras a produtos similares no mercado e aumentar o valor da empresa, bem como a possibilidade de obter mais faturamento com a licença.

Um outro entendimento sobre a patente é que ela é um contrato entre o inventor e o público, pelo qual o inventor, considerando que o uso exclusivo de sua invenção lhe é assegurado por um período limitado de tempo, confere ao público o conhecimento da invenção durante esse período, bem como um direito irrestrito de usá-la após esse período ter expirado (BIAGIOLI, 2006).

A patente pode ser utilizada como forma de garantir uma medida exata da atividade inovadora. Embora o patenteamento impeça que qualquer outro concorrente produza um produto idêntico ou use um processo idêntico, sem prévia autorização, ele deixa em aberto a opção de o concorrente produzir um produto diferenciado e usar um processo relacionado ou original (HORSTMANN *et al.*, 1985). A atividade moderna de patenteamento, além de objetivar a garantia do direito de uso exclusivo da marca e do processo de obtenção do produto pelo inventor, é significativamente impulsionada por motivo comercial, ou seja, obtenção de lucro por meio do monopólio da exploração da patente negociada. Do ponto de vista de pesquisadores, o patenteamento é motivado pela participação no desenvolvimento do processo de padronização, a reputação, o aumento de recursos e o fornecimento de fundos para pesquisas futuras (BLIND; POHLISCH; ZI, 2018).

O surgimento da ideia de patenteamento conjugado com o crescente volume de documentos de patentes, os litígios de documentos de patentes e a descoberta de que as invenções e as inovações, além de garantirem o direito de uso exclusivo da marca e do processo para a obtenção do produto, geram também valor comercial. Trazem também a necessidade de efetuar uma análise rigorosa de informações contidas em patente utilizando

ferramentas computacionais automatizadas. Esta análise rigorosa de informações da patente visando descobrir elementos importantes, suas inter-relações, assim como as tendências, é realizada pela *Patinformatics* ou Informática de Patentes ou ainda Patinformática (ABBAS; ZHANG; KHAN, 2014).

Entre as tarefas comumente conduzidas pela *Patinformatics* destacam-se: a limpeza e agrupamento de lista (do inglês *List cleanup and grouping*), onde a limpeza de lista é definida como a padronização manual ou automática de termos dentro de um campo de dados e o agrupamento, que também inclui a padronização, inclui a identificação de sinônimos e termos alternativos para descrever um conceito semelhante; e a Análise de Citações, onde são utilizadas árvores hiperbólicas para mostrar relações entre patentes que citam umas às outras (KALYAN; SANGEETHA, 2020; TRIPPE, 2003).

Shalaby (2019), apresentou tarefas típicas de análise de patentes. Entre elas destacam-se: a) a exploração de tecnologias visando capturar tecnologias novas e modernas em um domínio específico e usá-las posteriormente para criar serviços inovadores; b) a análise do cenário de tecnologia, visando avaliar a densidade de registros de patentes de tecnologia e direcionar atividades de P&D adequadamente; c) a análise competitiva e *benchmarking*, visando identificar pontos fortes e diferenças do portfólio de patentes de uma empresa em comparação com outros competidores que trabalham em tecnologias relacionadas; d) a classificação e pontuação de patentes (do inglês *patent ranking and scoring*), visando quantificar a força de reivindicações de uma patente existente ou nova; e e) a pesquisa de estado da arte, visando recuperar documentos de patentes e publicações científicas relevantes para um novo pedido de patente. Além disso, a tarefa de recuperação de patentes, uma das subáreas da Recuperação de Informação (RI), é considerada o pilar de quase todas as tarefas de análise de patentes. Citam-se ainda outras tarefas relacionadas à recuperação de patentes, entre elas, a avaliação de patentes, litígios e licenciamento (SHALABY; ZADROZNY, 2019).

Outra tarefa importante na análise de patentes é a recomendação de patentes a serem licenciadas, embora não existam muitas pesquisas focando na recomendação automática de potenciais licenças. Chen *et al.* (2009), por exemplo, propuseram uma plataforma ou sistema chamado SIMPLE que usa similaridade do vizinho mais próximo para encontrar outras patentes que são mais semelhantes ao conjunto de tópicos fornecido.

No contexto da recomendação de patentes existem diversos desafios. Entre eles destacam-se: a alta dimensionalidade dos textos, visto que uma patente é um documento extenso, as terminologias altamente complexas e de domínio específico e, por fim, a falta de padronização no estilo de redação de patentes (SHALABY; ZADROZNY, 2019). Somam-se a estes, os diferentes formatos em que as patentes se encontram, a capacidade dos especialistas ou examinadores em analisar patentes e a necessidade de cada examinador ao longo de tempo em adquirir diferentes habilidades para realizar uma avaliação adequada (ABBAS; SHANG; KHAN, 2014). Desta forma, há um trabalho importante em definir adequadamente o perfil de cada examinador de maneira que, no momento de avaliação da patente, seja escolhido o examinador ou especialista cujo perfil esteja mais aderente para a referida tarefa.

Para melhor caracterizar a relevância de se definir adequadamente o perfil dos examinadores, menciona-se que o número de pedidos de patentes teve um aumento

expressivo nas últimas décadas, passando de de 1.703.200 em 2005 para 3.224.200 em 2019 (BURKHARDT; DILEXIT; ANDERSSON, 2021). Este enorme volume de solicitações de patentes que chegam às autoridades avaliadoras e a crescente necessidade de acelerar o processo de análise de documentos e concessão de patentes, torna necessária a automatização do processo de escolha de profissionais especializados para atuarem na tarefa de análise de patentes.

Neste trabalho, propõe-se um método que, no contexto da análise de patentes, forneça, a partir de uma patente de interesse (entrada), a recomendação de um *ranking* de avaliadores com base no perfil dos mesmos. Para a implantação do método é proposta a utilização de *embeddings* como técnica principal. Já a construção e o treinamento do modelo de recomendação é realizado utilizando o algoritmo não supervisionado *Doc2Vec*. Para encontrar a similaridade entre determinada patente de entrada a ser avaliada e outras já avaliadas e, a partir disso, recomendar um *ranking* de possíveis examinadores, utiliza-se a medida de similaridade do cosseno.

Além desta seção, o trabalho está organizado em outras 5 (cinco) seções. A seção 2 apresenta os conceitos básicos e essenciais utilizados no desenvolvimento do trabalho. Na seção 3 são apresentados trabalhos correlatos à pesquisa no contexto de sistemas de recomendação de patentes. A seção 4 detalha o método proposto. Na sequência, é apresentada a análise e discussão dos resultados (seção 5) e, por fim, na seção 6, é apresentada a conclusão e trabalhos futuros.

## 2 FUNDAMENTAÇÃO TEÓRICA

### 2.1 ANÁLISE DE PATENTES

A vasta quantidade de patentes existentes e o seu contínuo crescimento têm chamado atenção sobre a necessidade de desenvolvimento de técnicas, ferramentas e metodologias sofisticadas para a análise de informações contidas em patentes visando revelar padrões nessas coleções (SJÖGREN *et al.*, 2020).

Atualmente, há várias técnicas utilizadas para a análise de patentes. Trappey *et al.* (2019), por exemplo, aplicaram o método analítico de aprendizado profundo para estimativa automática do valor da patente. A análise de componentes principais (do inglês *Principal component analysis - PCA*) é adotada para a seleção automática de indicadores de patentes antes da aplicação de modelos de redes neurais profundas para a estimativa de valor de patente.

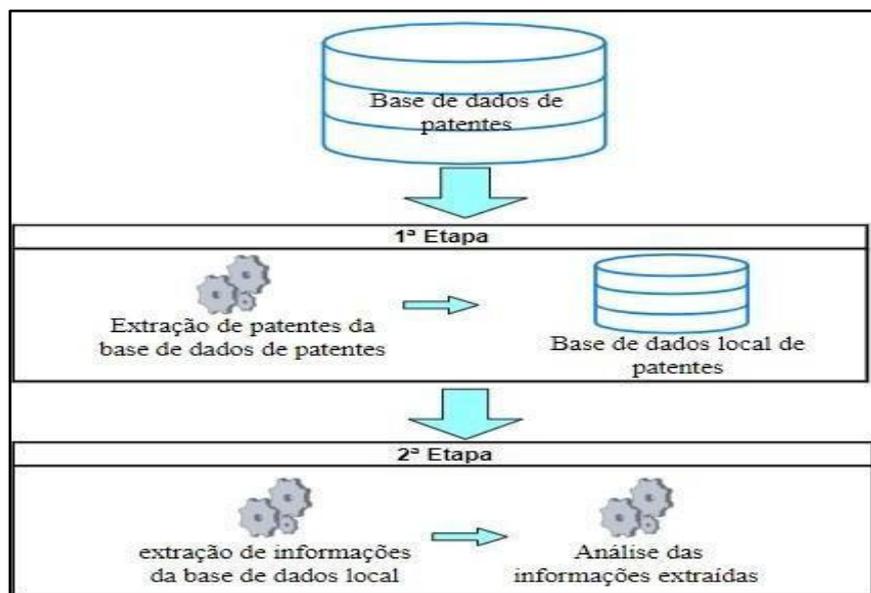
Abbas *et al.* (2014), apresentaram cinco técnicas utilizadas para a análise de patentes, sendo, a mineração de textos, a visualização, a análise de citações, e a geração de mapas de patentes e mapas de tecnologia (do inglês *Technology Road-Mapping - TRM*). Estas técnicas auxiliam especialistas em análise de patentes na realização de tarefas de previsão de tendências tecnológicas, previsão do desenvolvimento tecnológico em um domínio particular, detecção de violações de patentes, extração de informações de patentes para identificar infrações, determinação de qualidade das patentes, as patentes mais promissoras e vazios de patentes.

A Análise do Panorama de Patentes (do inglês *Patent Landscape Analysis - PLA*) é uma outra técnica de análise de patentes utilizada pelos *experts* em análise de patentes. É definida como uma pesquisa do estado da arte de patentes de última geração e fornece representações gráficas que demonstram tendências de patentes, principais cessionários de patentes, parceiros de colaboração e outros detalhes. O principal foco desta técnica são as patentes e pedidos de patentes de uma determinada área de tecnologia ou portfólio de patentes de uma empresa (YANG *et al.*, 2010).

A análise de informações de patentes envolve uma série de etapas (Figura 1). Entre elas, destacam-se a extração de patentes de bancos de dados de patentes, a extração de informações de patentes e a análise das informações extraídas para inferir conclusões lógicas (ABBAS; ZHANG; KHAN, 2014).

Existem diferentes níveis de análise de informações de patentes. Trippe (2003) propôs três níveis macro de análise de informações de patentes no contexto da Patinformática, sendo: a) Inteligência de Patentes, na qual as informações desses textos são usadas para identificar as capacidades técnicas de uma organização e o uso dessa inteligência para desenvolver estratégias de planejamento técnico estratégico; b) Mapeamento de Patentes, que usa dados de patentes publicadas para criar representações gráficas ou físicas de arte relevante pertencente a uma área específica ou nova invenção e; c) Análise de Citação de Patentes, relacionada ao estudo de citações de patentes para determinar o valor de uma patente, a identificação de potenciais parceiros de licenciamento ou *leads* com base na citação de patentes de uma organização por outra empresa ou espaço de mercado completamente diferente.

Figura 1 - Etapas genéricas da análise de patentes



Fonte: Adaptado de Abbas, Zhang e Khan (2014).

### 2.1.1 Tarefas de Análise de Patentes

O processo de análise de informações de patentes envolve inúmeras tarefas. Segundo Trippe (2003), existem nove tarefas comumente conduzidas pela Patinformática, sendo 6 (seis) de maior destaque:

- Limpeza e agrupamento de dados (coleção ordenada de valores): A limpeza refere-se a padronização manual ou automática de termos dentro de um campo de dados. O agrupamento, por sua vez, inclui a padronização, como nos casos em que os termos com erros ortográficos são limpos, e inclui também a identificação de sinônimos e termos alternativos para descrever conceitos semelhantes. Além disso, permite que termos sinônimos sejam combinados para que seus significados em um conjunto de dados possam ser avaliados com precisão.
- Análise de Citações: são utilizadas árvores hiperbólicas para mostrar relações entre patentes que citam umas às outras. Contagens de citações para encontrar documentos potencialmente essenciais também fazem parte desta tarefa.
- Agrupamento de Dados Estruturados: é a organização de grandes volumes de documentos, em que documentos que compartilham conteúdos em comum são atribuídos ao mesmo grupo. Nesta tarefa são considerados os metadados das patentes.
- Agrupamento de Dados Não Estruturados: similar a tarefa anterior, mas se utiliza de processamento do texto para identificar conceitos e frases, para então, identificar documentos similares de modo que estes pertençam a determinado grupo.
- Mapeamento de Grupos de Documentos: é uma tarefa na qual são organizados grupos de documentos em um espaço bidimensional ou tridimensional, sendo que coleções de documentos que compartilham elementos em comum são dispostas de maneira mais próxima. Este tipo de mapeamento permite a identificação de relação entre coleções de documentos, possivelmente, de diferentes áreas de conhecimento.
- Adição de Componentes Temporais à Mapas de Agrupamentos: possibilita que informações temporais sejam exibidas em cores diferentes de modo que a progressão de um item ao longo do tempo seja rastreada.

Já Krestel *et al.* (2021), apresentam 8 tarefas de análise de patentes mais populares identificadas na literatura, sendo: a) Tarefa de suporte, que envolve pré-processamento, extração de informações para análise de patentes ou tradução de patentes para outros idiomas; b) Classificação de patentes, onde os documentos de patentes são categorizados hierarquicamente com base no campo de invenção; c) Recuperação de patentes, onde são desenvolvidas as técnicas e métodos de recuperação de documentos de patentes relevantes considerando uma determinada pesquisa; d) Avaliação de patentes e previsão do valor de mercado, onde o conteúdo e os detalhes bibliográficos das patentes são analisados com certos protocolos para analisar a qualidade dos pedidos de patentes; e) Previsão de tecnologia, onde as patentes são usadas para avaliar um cenário tecnológico com o intuito de auxiliar pesquisadores na identificação de novas tecnologias; f) Geração de texto de patente, onde a estrutura e os estilos incorporados aos documentos de patentes publicados são usados para automatizar o processo de redação de reivindicações de patentes; g) Análise

de litígios, visando identificar potenciais patentes que estejam em disputa ou litígio entre duas empresas ao proibir o desenvolvimento de estratégias de negócios; e h) Tarefa de visão computacional, visando auxiliar no entendimento com figuras e desenhos que constam em documentos de patentes.

### 2.1.2 Desafios de Análise de Patentes

Com o volume de informações sobre patentes cada vez maiores, as tarefas de análise de patentes tornaram-se vitais tanto do ponto de vista jurídico, administrativo quanto de gestão de patentes. Consequentemente, os dados de patentes são analisados de diferentes formas para atender a diferentes propósitos. No entanto, analisar patentes manualmente constitui um desafio árduo e demorado para grandes volumes de dados textuais, além de desafios relacionados à capacidade de especialistas em análise de patentes. Portanto, o processo de análise de patentes requer procedimentos e ferramentas automatizadas bem definidas para diminuir as dificuldades enfrentadas pelos analistas de patentes na pesquisa em grandes volumes de documentos de patentes (ABBAS; ZHANG; KHAN, 2014; TANG *et al.*, 2012).

Um outro exemplo do desafio de análise de patentes se refere à urgência na análise de grandes quantidades de documentos de patentes relacionadas ao tratamento de águas residuais industriais. Para este caso em particular, Mao *et al.* (2022) propuseram a análise de patentes combinada com mineração de texto para analisar quantitativamente 11.840 patentes relacionadas ao tratamento de águas residuais industriais no banco de dados *Derwent Innovations Index*.

## 2.2 SISTEMAS DE RECOMENDAÇÃO

Os diversos desafios enfrentados por especialistas no que tange às diversas tarefas de análise de patentes levando em conta o grande volume de documentos neste tipo de cenário, constitui uma preocupação com foco na automatização dessas tarefas. Para tal, a adoção de técnicas e métodos providas pela área de Sistema de Recomendação, pode auxiliar na diminuição dessas dificuldades enfrentadas no processo de análise de patentes.

Hawashin *et al.* (2019) e Cha *et al.* (2019), definem Sistemas de Recomendação (do inglês *Recommender System - RS*) como sistemas que promovem sugestões aos usuários com base em seus interesses. Entre outras aplicações de RSs têm-se a recomendações de produtos de lojas *online*, recomendações de amigos, recomendações de livros e muito mais. Neste trabalho, o foco reside na recomendação de examinadores para a avaliação de patentes, em que o perfil dos examinadores é definido de acordo com as patentes previamente avaliadas.

Os RSs podem ser definidos também como sistemas ou programas que procuram recomendar itens mais adequados (produtos ou serviços) para usuários específicos (indivíduos ou organizações), prevendo o interesse de um usuário em um item com base nas informações relacionadas sobre os itens, os usuários e as interações entre itens e usuários (LÜ *et al.*, 2015).

Para LÜ *et al.* (2012), os RSs auxiliam na entrega de conteúdos certos às pessoas certas. Para tal, torna-se necessária a utilização de diferentes abordagens. LÜ *et al.* (2015) e Aggarwal (2016), apresentaram oito e cinco abordagens de RSs, respectivamente. A seguir, são apresentadas as quatro abordagens mais comuns:

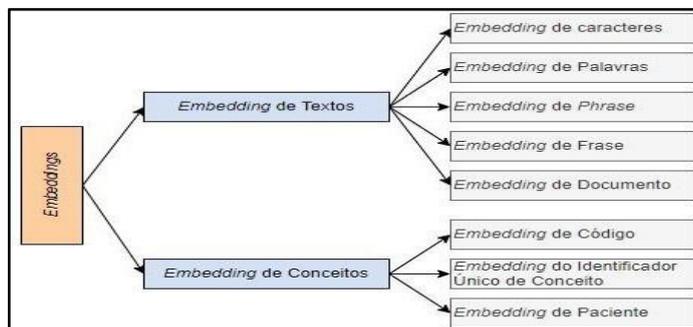
- **Recomendação Baseada em Conteúdo** (do inglês *Content-Based Recommender System - CRS*): os atributos descritivos dos itens são usados para fazer recomendações. Nesta abordagem, as avaliações e o comportamento dos usuários são combinados com as informações de conteúdo disponíveis nos itens para realizar recomendações.
- **Recomendação Baseada em Filtragem Colaborativa** (do inglês *Collaborative Filtering - CF*): utilizam informações colaborativas geradas pelos usuários para produzir recomendações. Esta abordagem possui alguns desafios, destacando-se a esparsidade das matrizes de classificação subjacentes e o problema da partida fria, ou seja, a falta de dados em um primeiro momento.
- **Recomendação Baseada em Conhecimento** (do inglês *Knowledge-Based Recommender Systems - KRS*): fornece recomendações com base na especificação explícita do usuário do tipo de conteúdo que este deseja. Esta abordagem é útil, por exemplo, no contexto de itens que não são comprados com muita frequência.
- **Sistemas de Recomendação Híbridos e Compostos** (do inglês *Hybrid and Ensemble-Based Recommender Systems - HERS*): *hybrid* refere-se à combinação de vários aspectos de diferentes abordagens de RSs para alcançar melhores resultados na resolução de determinada tarefa de recomendação e, *ensemble*, combina o poder de múltiplas fontes de dados e vários modelos para melhorar a eficácia dos sistemas de recomendação e obter recomendações mais precisas.

## 2.3 EMBEDDINGS

Os *Embeddings*, alternativamente chamados de representações vetoriais distribuídas (LE; MIKOLOV, 2014), referem-se à técnica de representação numérica (mapeamento) de textos de comprimento variável para vetores de comprimento fixo ou densos, mantendo as informações mais relevantes. Diferem de representações baseadas na frequência da palavra como *tf-idf* (*term-frequency/inverse document frequency*) e *n-grams* que ignoram a ordem, assim como as semelhanças sintáticas e semânticas das palavras (KALYAN; SANGEETHA, 2020; GARTEN *et al.*, 2015).

CHEN *et al.* (2018) propuseram dois modelos de *embeddings*, de palavras (do inglês *Word Embeddings - WEs*) e de documentos (do inglês *Document Embeddings - DEs*). Por outro lado, KALYAN e SANGEETHA (2020), propuseram a classificação dos *embeddings* em dois grandes grupos (Figura 2): *Embeddings* de Texto e *Embeddings* de Conceito. O primeiro é classificado em *Character Embedding*, *Word Embedding*, *Phrase Embedding*, *Sentence Embedding* e *Document Embedding*. O *Embedding* de Conceito é classificado em *Embedding* de Código, *Embedding* do Identificador de Conceito e *Embedding* de Paciente.

Figura 2 - Classificação de Embeddings



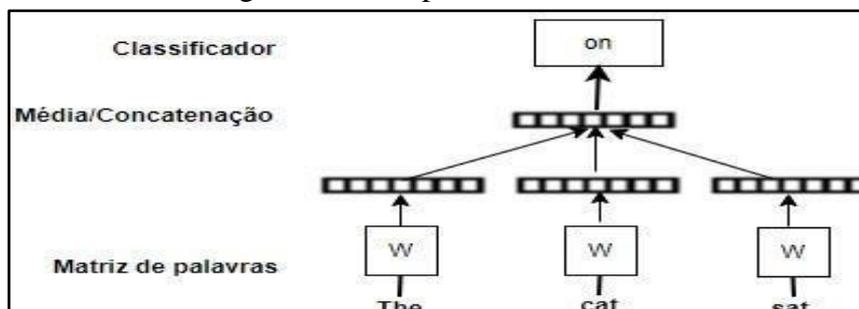
Fonte: Adaptado de Kalyan e Sangeetha (2020).

### 2.3.1 Métodos de *Word Embeddings*

*Word Embeddings* são vetores de palavras densos, distribuídos, de comprimento fixo, construídos usando estatísticas de coocorrência de palavras de acordo com a hipótese de distribuição (ALMEIDA; XEXÉO, 2019) e se concentram mais na relação das palavras (CHEN *et al.*, 2018). Na literatura, segue alguns métodos frequentemente utilizados para a geração do conceito de embeddings (DUMAIS, 2004; KALYAN; SANGEETHA, 2020):

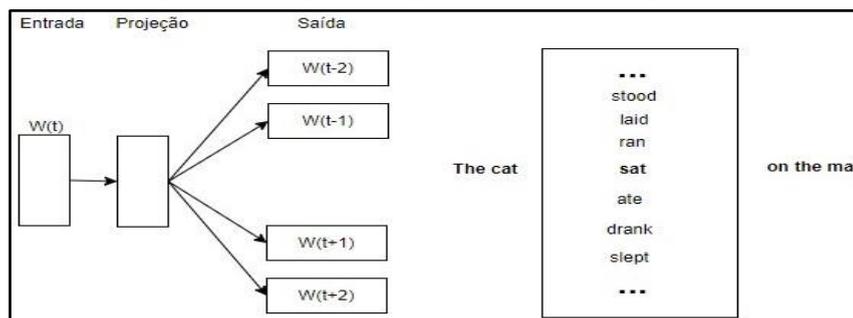
- *Latent Semantic Analysis (LSA)*: é uma abordagem que modela simultaneamente as relações entre os documentos com base em suas palavras constituintes e as relações entre as palavras com base em sua ocorrência nos documentos. Não faz uso de técnicas de processamento de linguagem natural para analisar relações morfológicas, sintáticas ou semânticas. Inicia com uma grande coleção de textos, constrói uma matriz de termos e de documento e procura descobrir algumas estruturas de similaridade que são úteis para a recuperação de informações e problemas de análise de texto relacionados (DUMAIS, 2004).
- *Word2Vec*: é representado por dois modelos de redes neurais: o *Continuous Bag-Of-Words (CBOW)*, Figura 3 e o *Skip-Gram*, Figura 4. Ele é utilizado para representações distribuídas de palavras em um *corpus*. Esses modelos possuem uma camada de entrada, uma camada oculta e uma camada de saída.

Figura 3 - Exemplo do modelo CBOW



Fonte: Adaptado de Le e Mikolov (2014).

Figura 4 - Representação do modelo Skip-Gram



Fonte: Adaptado de Yue e Li (2020), TASNIM *et al.* (2021).

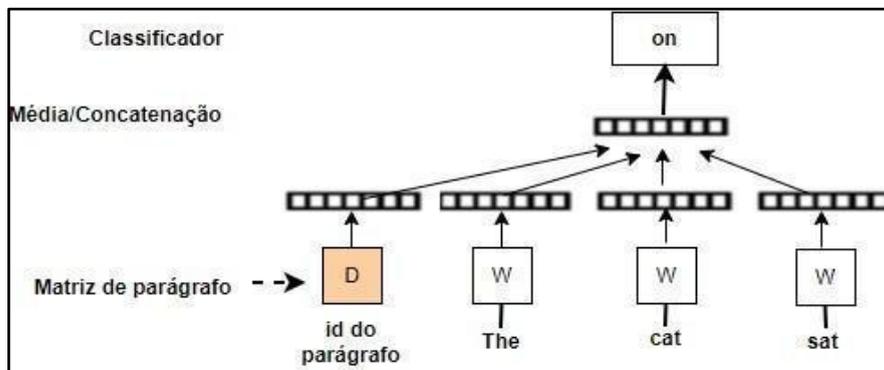
- *GloVe* (*Global Vectors*): diferente do *Word2Vec* que incorpora apenas estatísticas locais (informações de contexto local de palavras), *GloVe* incorpora estatísticas globais (coocorrência de palavras) e estatísticas locais para obter vetores de palavras. A desvantagem desse modelo é que ele ignora informações morfológicas, bem como a natureza polissêmica das palavras (KALYAN; SANGEETHA, 2020). O modelo *GloVe* tenta gerar a representação vetorial da palavra usando a similaridade entre as palavras como um invariante. O modelo explora abordagens fornecidas pelos dois modelos do *Word2Vec*, *Skip-Gram* e *Continuous Bag of Words* (SHARMA *et al.*, 2017).
- *FastText*: é um modelo que codifica informações morfológicas em vetores de palavras. Diferentemente de *Word2Vec* e *GloVe*, que tratam palavras como unidades atômicas e atribuem representações vetoriais ignorando informações de subpalavra e sofrendo com problemas de OOV (do inglês *out-of-vocabulary*, representando termos que não fazem parte do dicionário léxico normal do método), *FastText* é a modificação do modelo *Skip-Gram* para alavancar informações de subpalavra e fornecer vetores para palavras raras e OOV.
- *ELMo* (*Embeddings from Language Models*): é um modelo que gera *embeddings* (representações vetoriais densas) para uma palavra considerando seu contexto, fazendo uso de embeddings de caracteres e BiLSTM (*Bidirectional Long Short Term Memory*). Diferente de *Word2Vec*, *GloVe* e *FastText* que atribuem uma única representação vetorial a uma palavra independente do contexto em que é usada, *ELMo* captura o contexto da palavra e, portanto, representa de maneira mais adequada a natureza polissêmica das palavras.

### 2.3.2 Métodos de *Document Embedding*

*Document Embeddings* referem-se à representação numérica vetorial de comprimento fixo (denso) de textos de comprimento variável, como sentenças, parágrafos e documentos (LE; MIKOLOV, 2014). Na literatura, o método de *document embedding* frequentemente utilizado é o *Paragraph2vector*, popularmente conhecido como *Doc2vec*. *Doc2Vec* é uma extensão do *Word2Vec*, sendo um modelo não supervisionado que mapeia texto de comprimento variável como frases, parágrafos e documentos para representações vetoriais densas. *Doc2vec* oferece dois modelos: *Distributed Memory version of Paragraph Vector* (*PV-DM*) e *Distributed Bag of Words version of Paragraph Vector* (*PV-DBOW*) (KALYAN; SANGEETHA, 2020).

- PV-DM*: é semelhante ao modelo *Continuous Bag of Words (CBOW)* do *Word2Vec*. No entanto, enquanto *CBOW* prevê a palavra central a partir de palavras de contexto, *PV-DM* prevê a próxima palavra usando a concatenação ou média dos vetores de parágrafo e palavras de contexto. É composto de três camadas. A primeira camada utiliza vetores de palavras de parágrafo e contexto como entrada. Já a segunda camada concatena ou calcula a média de ambos os vetores. Por fim, a camada final, representada por um classificador, prevê o vetor para a próxima palavra (KALYAN; SANGEETHA, 2020). Neste modelo, cada parágrafo é mapeado para um vetor único, representado por uma coluna em uma matriz  $D$  e cada palavra também é mapeada para um vetor único, representado por uma coluna em uma matriz  $W$ . O vetor de parágrafo e os vetores de palavras são calculados em média ou concatenados para prever a próxima palavra em um contexto. Considerando o exemplo que consta na Figura 5, a próxima palavra prevista é “on”.

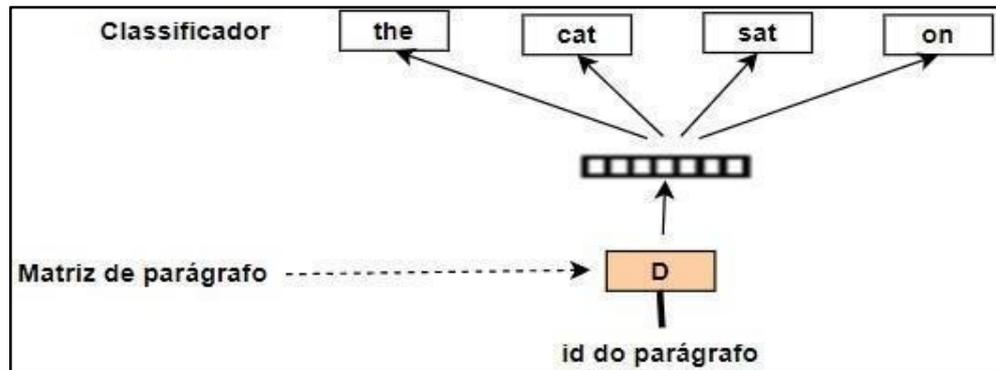
Figura 5 - Exemplificação do modelo *PV-DM*



Fonte: Adaptado de Le e Mikolov (2014).

- PV-DBOW*: este modelo é semelhante ao modelo *Skip-Gram* do *Word2Vec*. Mas, enquanto *Skip-Gram* prevê palavras de contexto a partir da palavra central/atuai, *PV-DBOW* prevê palavras de contexto usando parágrafos. *PV-DBOW* consiste de três camadas, sendo que a primeira camada utiliza o vetor do parágrafo como entrada. Já a segunda camada é a camada oculta e, por fim, a camada final que representa um classificador responsável por prever os vetores das palavras do contexto (KALYAN; SANGEETHA, 2020). O modelo *PV-DBOW* considera a concatenação do vetor de parágrafo com os vetores de palavras para prever a próxima palavra em uma janela de texto. Na Figura 6 é apresentado o modelo *PV-DBOW* de forma simplificada.

Figura 6 - Exemplificação do modelo PV-DBOW



Fonte: Adaptado de Kalyan e Sangeetha (2020).

### 3 TRABALHOS CORRELATOS

Através de buscas na literatura científica foram selecionados trabalhos a fim de identificar métodos e abordagens utilizadas na construção de sistemas de recomendação de examinadores para avaliação de patentes. A pesquisa foi efetuada em bases de dados como Scopus®, Web of Science®, ScienceDirect®, IEEE Xplore®, ACM Digital Library® e Springer Link®. Num primeiro momento utilizou-se a string de busca ("Recommender System" OR "Recommender Systems") AND "Patent Recommendation" AND (Expert OR Experts OR Examiner OR Examiners), sendo os resultados pouco representativos. A maioria das bases não retornou artigos relacionados. Neste sentido, as palavras Expert e Examiner, assim como o plural destas palavras, foram retiradas da *string* de busca. Uma nova busca utilizando a *string* ("Recommender System" OR "Recommender Systems") AND "Patent Recommendation" permitiu selecionar trabalhos relacionados a fim de identificar métodos e abordagens utilizadas na construção de sistemas de recomendação no cenário de avaliação de patentes. O Quadro 1 ilustra as bases de dados em que as pesquisas foram realizadas e as respectivas quantidades de trabalhos.

Quadro 1 – Bases de dados em que as pesquisas foram realizadas e as respectivas quantidades de trabalhos.

Base de dados	Quantidade de trabalhos recuperados
Scopus®	8
Web of Science®	3
Science Direct®	3
IEEE Xplore®	4
ACM Digital Library®	0
Springer Link®	3

Fonte: Elaborada pelo autor (2022).

Dos trabalhos recuperados após as buscas na literatura científica, seis foram selecionados e utilizados como base para o desenvolvimento da presente seção, considerando como critério de seleção os trabalhos publicados a partir do ano de 2010. A

seguir estão descritos os estudos selecionados que possuem relação, em alguma medida, com o contexto desta pesquisa.

Trappey *et al.* (2013) desenvolveram um sistema de recomendação que visa evitar a potencial violação dos Direitos de Propriedade Intelectual (DPI) e garantir novos DPI. O sistema desenvolvido agrupa usuários com comportamentos de pesquisa de patentes semelhantes e, na sequência, infere novas recomendações de patentes com base nos comportamentos e características dos membros do grupo. Como abordagem, propuseram a Filtragem Colaborativa que permite a pesquisa dinâmica de patentes relacionadas em bancos de dados globais de patentes. Foram usadas 987 patentes relacionadas da base de dados da USPTO<sup>®</sup> para construir o banco de dados de patentes que inclui os registros de comportamento dos vinte pesquisadores que conduziram as operações de busca de patentes.

Com o objetivo de auxiliar pesquisadores interessados em publicar seus artigos ou papers no processo de escolha de periódico adequado para a publicação de seus trabalhos, o que ajudaria a diminuir as submissões incorretas que acabam resultando na perda de tempo e esforço tanto da parte dos pesquisadores quanto da parte dos editores, Jain *et al.* (2019) propuseram o Sistema de Recomendação de Periódicos (do inglês *Journal Recommendation System - JRS*). Para o desenvolvimento do sistema foi utilizado a abordagem de filtragem colaborativa baseada em conteúdo (do inglês *Content-based filtering*). Dado que o sistema proposto visa recomendar periódicos relevantes, é necessário encontrar similaridade entre os detalhes do artigo de entrada fornecidos pelo usuário e os detalhes do periódico no conjunto de dados, sendo utilizada a distância euclidiana para esse fim. O conjunto de dados utilizado neste trabalho é preparado pelos autores.

Em seu trabalho intitulado “*embedding* de rede de informações heterogêneas com base semântica para recomendação de citação de patente” (do inglês *Semantic Based Heterogeneous Information Network Embedding for Patent Citation Recommendation*), Zhang *et al.* (2020), apresentaram um método que obtém as relações semânticas latentes entre patentes por similaridade de conteúdo textual e similaridade de tópicos, e liga as relações semânticas com a estrutura de rede para construir uma nova rede de informação heterogênea. Para combinar as duas similaridades, utilizaram o método de fusão linear. O modelo proposto foi avaliado utilizando 153.777 patentes da base de dados da USPTO<sup>®</sup>.

Bai *et al.* (2021) propuseram o método chamado *Collaborative Filtering Auto-Encoder* para a recomendação de *top-N* patentes. O método visa rastrear os interesses dos usuários e melhorar a qualidade da recomendação quando a matriz de classificação é extremamente esparsa. Para isso, o método emprega a abordagem *Auto-Encoders* para extrair os recursos do item, converter um vetor esparsa de alta dimensão em um vetor denso de baixa dimensão e, em seguida, usar o vetor denso para o cálculo de similaridade. Para o desenvolvimento deste trabalho foram utilizados os conjuntos de dados MovieTweetings<sup>®</sup>, hetrec2011<sup>®</sup>, learning-from-sets-2019<sup>®</sup> e CiaoDVD<sup>®</sup>, apesar destes conjuntos não estarem relacionados ao contexto de patentes.

Trappey *et al.* (2021) desenvolveram uma metodologia para um recomendador de patentes descobrir patentes semanticamente relevantes para mineração de tecnologia e análise de tendências. O recomendador de patentes proposto adota algoritmos de aprendizado de máquina para processamento de linguagem natural com o intuito de representar documentos de patentes em espaço vetorial e, assim, permitir análises semânticas dos documentos de patentes. O método de aprendizado de máquina utilizado

para treinar o modelo de rede neural é o *Doc2Vec* e as bases de dados utilizados para este trabalho são o USPTO<sup>®</sup>, o *European Patent Office (EPO)*<sup>®</sup>, o *China's State Intellectual Property Office (SIPO)*<sup>®</sup> e *World Intellectual Property Organization (WIPO)*<sup>®</sup>.

Embora a grande maioria dos trabalhos encontrados na literatura ao longo da pesquisa realizada sejam relacionados a sistemas de recomendação de patentes, o trabalho que mais se aproxima ao objetivo deste trabalho voltado à recomendação de avaliadores para a análise de patentes é o artigo de FU *et al.* (2015). Os autores desenvolveram um estudo que visa construir um sistema automático de recomendação de citações de patentes para examinadores de patentes na USPTO<sup>®</sup>. O método utilizado na construção do sistema automático de recomendação de citações de patentes para examinadores de patentes é o *Support Vector Machine for Ranking (RankSVM)* e *Least Angle Regression (LAR)*. O sistema proposto foi avaliado usando um conjunto de dados reais de mais de 2,4 milhões de patentes dos EUA emitidas de 1995 a 2010.

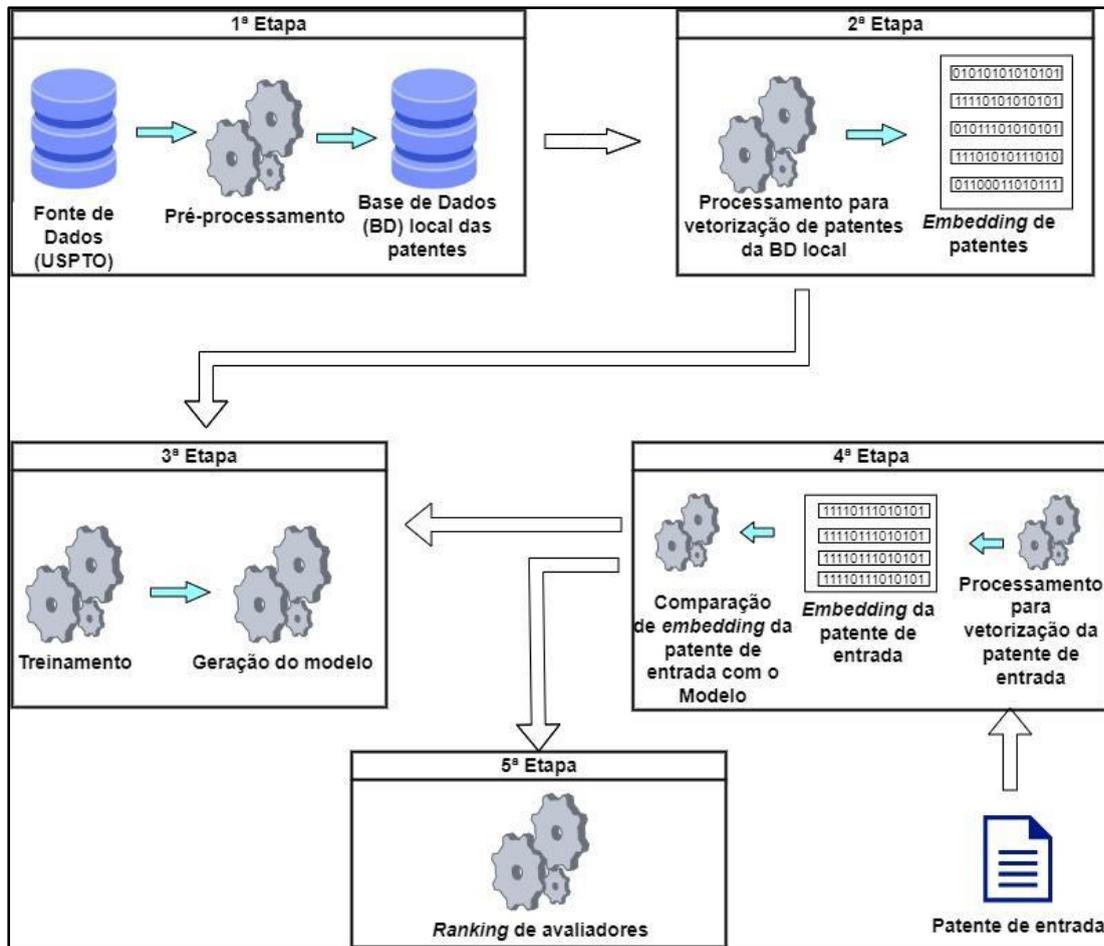
Analisando os trabalhos apresentados percebe-se o uso de abordagens como a de filtragem colaborativa baseada no comportamento dos usuários, filtragem colaborativa baseada em conteúdo, a fusão linear, o **Auto-Encoders**, assim como pode-se verificar também no trabalho de Trappey *et al.* (2021) o uso de *Doc2Vec* como técnica para realizar as recomendações das patentes. A técnica *Doc2Vec*, juntamente com o conceito de *embedding*, foram consideradas também neste trabalho. Além disso, verifica-se uma clara tendência nos trabalhos analisados para a recomendação de citações de patentes, enquanto que a presente pesquisa se concentra na recomendação de examinadores para a análise de patentes.

#### 4 MÉTODO PROPOSTO

O método proposto neste trabalho é fundamentado na análise de dados não estruturados representados por patentes e na vinculação dessas patentes aos examinadores. Para tal, são utilizados métodos de representação vetorial na forma de vetores densos (*embeddings*), possibilitando que, a partir de uma determinada demanda, isto é, a apresentação de uma patente, seja recomendado um conjunto ordenado de especialistas/examinadores. Além disso, o método é suportado pela abordagem de recomendação baseada em conteúdo.

De maneira geral, o método é constituído de cinco etapas principais (Figura 7). A primeira etapa refere-se à coleta e pré-processamento de documentos das patentes; a segunda etapa trata da transformação de documento de patentes na forma de vetores densos por meio do conceito de *embedding*; a terceira etapa representa o treinamento e geração do modelo de recomendação; a quarta etapa refere-se ao processamento para a transformação de uma patente de entrada em um vetor denso (*embedding*) possibilitando a consulta no modelo gerado na etapa anterior conduzindo então, a quinta e última etapa, sendo esta responsável por recomendar um *ranking* de potenciais examinadores para a patente de interesse.

Figura 7 - Visão geral das cinco etapas do modelo proposto



Fonte: Elaborado pelo autor (2022).

#### 4.1 ETAPA 1: COLETA E PROCESSAMENTO DE DOCUMENTOS DAS PATENTES

Nesta primeira etapa é efetuada a coleta de documentos de patentes disponibilizados na *web* de maneira aberta. Por questões de facilidade de obtenção dos dados, a base de patentes da *United States Patent and Trademark Office*<sup>®</sup> (USPTO) foi considerada. A USPTO<sup>®</sup> é uma agência federal para concessão de patentes e registro de marcas dos Estados Unidos da América. Todavia, qualquer base de patentes poderia ser coletada.

Na sequência, são identificados os principais campos com conteúdo textual que possam melhor representar determinada patente. A estes campos aplica-se um conjunto de operações de limpeza, tais como, retirada de pontuações e de palavras não relevantes (*stop words*), caracterizando-se como pré-processamento. Além disso, para atender a proposta do método, é requerida a identificação e extração do examinador de determinada patente.

De modo a facilitar a manipulação das patentes e evitar o reprocessamento, o método prevê o armazenamento em uma base de dados considerando os principais campos, por exemplo, título, resumo e o examinador. Isto também se justifica pelo volume de dados para este tipo de conteúdo.

#### 4.2 ETAPA 2: TRANSFORMAÇÃO DE DOCUMENTO DE PATENTES EM *EMBEDDINGS*

Nesta etapa, as patentes da base de dados local são processadas e transformadas em vetores densos, chamados de *embeddings*, ou seja, os documentos de patentes são mapeados para representações vetoriais com o intuito de preservar a semântica. Inicialmente ocorre a obtenção dos atributos preditores (conteúdo textual da patente a ser vetorizado) e o atributo meta (representado pelo nome do examinador) a partir da base de dados das patentes.

A partir disso, cada patente passa por um processo de tokenização, onde: i) são extraídos todos os *tokens* (palavras distintas) do conjunto de patentes formando um dicionário de  $n$  palavras; e ii) cada patente é transformada para uma representação vetorial chamada de *one-hot-encoding*, ou seja, um vetor  $n$  dimensional sendo que cada posição representa um *token* que será preenchido com 0 (zero) ou 1 (um) caso ocorra no texto da patente.

Por fim, considerando os vetores codificados de patente, aplica-se alguma técnica de geração de *embeddings*. Neste trabalho, o foco reside na utilização do conceito de *embeddings* de documentos.

#### 4.3 ETAPA 3: TREINAMENTO E GERAÇÃO DO MODELO

Após a geração dos *embeddings*, o conjunto de dados de patentes é dividido em dados de treinamento e dados de teste. Diversas abordagens podem ser utilizadas, assim como diferentes bibliotecas que facilitam este processo.

O conjunto de treinamento é utilizado para produzir o modelo de recomendação de examinadores, ou seja, o modelo que possibilitará a sugestão de um *ranking* de examinadores, do mais aderente para o menos aderente à determinada patente de entrada. Para o treinamento e geração do modelo, considerando a utilização do *framework Doc2vec*<sup>®</sup>, os seguintes parâmetros apresentados no Quadro 2 foram utilizados.

Quadro 2 - Definição dos parâmetros utilizados no *Doc2vec*<sup>®</sup>

<i>alpha</i>	0.025
<i>min_alpha</i>	0.00025
<i>min_count</i>	2
<i>dm</i>	1
<i>window</i>	2
<i>workers</i>	4
<i>vector_size</i>	200
<i>epochs</i>	100

Fonte: Elaborado pelo autor (2022).

A seguir consta a explicação de cada um dos parâmetros:

- *alpha*: é opcional e representa a taxa de aprendizado inicial;
- *min\_alpha*: é opcional e é o parâmetro que ao ser definido permite que a taxa de aprendizado caia linearmente para *min\_alpha* conforme o treinamento avance;
- *min\_count*: é opcional e é o parâmetro que ao ser definido permite ignorar todas as palavras com frequência total inferior a este valor;
- *dm*: define o algoritmo de treinamento. Se *dm=1*, *PV-DM* é usado. Caso contrário, *PV-DBOW* é usado;
- *window*: é opcional e define a distância máxima entre a palavra atual e a prevista em uma frase;
- *workers*: é opcional e representa o número de *threads* para treinar o modelo (o treinamento será impactado positivamente com máquinas *multicore*);
- *epochs*: é opcional e trata do número de iterações sobre o corpus. O padrão é 10 épocas para o *Doc2vec*<sup>®</sup>.

#### 4.4 ETAPA 4: GERAÇÃO DE *EMBEDDING* PARA UMA PATENTE DE ENTRADA

Esta etapa inicia com o provimento de um documento de patente. Este documento é então pré-processado através de um processo de tokenização (similar ao que ocorre na Etapa 2) gerando, a partir disso, um *embedding* da patente de entrada.

Na sequência, utilizando o *embedding* do documento da patente de entrada, é realizada uma consulta no modelo gerado na etapa anterior (Etapa 3) visando determinar quais são os documentos (patentes) mais similares. O modelo então retorna um conjunto de *n* documentos com as respectivas similaridades levando-se em conta o estabelecimento de uma similaridade entre a base de vetores e o vetor de entrada. Várias medidas de similaridade podem ser utilizadas, mas no contexto deste trabalho, a medida do cosseno é considerada devido a sua ampla utilização. A similaridade por cosseno é uma medida que computa a similaridade entre dois vetores e, no caso de vetores em que todas as dimensões possuem valores maiores ou iguais a 0 (zero), retorna valores entre 0 e 1, onde o valor 0 indica que não há similaridade entre os documentos e o valor 1 indica que os documentos são idênticos (KOCHER, SAVOY, 2017; PARK, HONG, KIM, 2020).

Assim, considerando dois documentos, *doc1* e *doc2*, o grau de similaridade entre eles pode ser expresso pela Equação 1 abaixo.

$$Sim(doc1, doc2) = \frac{doc1 \cdot doc2}{||doc1|| ||doc2||} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (1)$$

#### 4.5 ETAPA 5: RANKING DE AVALIADORES

A quinta e última etapa corresponde ao provimento de uma lista ordenada (*ranking*) de examinadores a partir de uma determinada patente de interesse.

Para tal, considerando o resultado obtido na etapa anterior (Etapa 4, na qual o modelo retorna  $n$  documentos com as respectivas similaridades) torna-se necessário determinar quais são os examinadores candidatos para avaliação de uma patente. Sendo assim, para cada uma das patentes pertencentes a lista verifica-se o examinador associado (Quadro 3).

Quadro 3 - Lista de  $n$  patentes com as respectivas similaridades e examinadores retornados a partir de uma consulta

<b>Identificação da Patente</b>	<b>Similaridade</b>	<b>Examinador</b>
Patente 01	0,87	Examinador 1
Patente 10	0,85	Examinador 2
Patente 55	0,79	Examinador 3
Patente 02	0,78	Examinador 1
Patente 88	0,75	Examinador 1
Patente 22	0,69	Examinador 3

Fonte: Elaborado pelo autor (2022).

Considerando as informações no quadro acima é criada uma nova lista agregada (Quadro 4), composta por examinadores e a quantidade de patentes avaliadas por cada um, sendo apresentados os examinadores considerando a ordem decrescente pela quantidade de patentes associadas conforme Quadro 3. Ou seja, a partir de uma lista de  $n$  patentes, a lista agregada terá, ao final, o examinador  $a$  com  $x$  documentos associados, o examinador  $b$  com  $y$  documentos associados, e assim por diante, em ordem decrescente pelo total de patentes associadas a cada examinador.

Quadro 4 - Lista de examinadores e respectivas patentes associadas

<b>Examinador</b>	<b>Quantidade de Patentes Associadas</b>
Examinador 1	3
Examinador 3	2
Examinador 2	1

Fonte: Elaborado pelo autor (2022).

## 5 ANÁLISE DOS RESULTADOS

### 5.1 DESCRIÇÃO DO CENÁRIO

O cenário de estudo deste trabalho envolve a recomendação de examinadores para determinada patente de interesse. Para tal, como resposta a determinada consulta a base de patentes é provida uma lista dos examinadores mais aderentes para realizar a análise, com o objetivo de facilitar a tomada de decisão dos gestores que possuem a função de atribuir uma patente a um examinador.

Neste sentido, de modo a permitir a avaliação deste cenário de estudo, a base de dados de patentes utilizada neste trabalho foi obtida a partir do trabalho de Silva (2022), sendo composto por patentes disponibilizadas pelo USPTO<sup>®</sup>.

Os dados disponibilizados são compostos por dois conjuntos de patentes e avaliadores, um com 16 avaliadores e 800 patentes associadas (12800 patentes ao todo) e outro com 26 avaliadores e 400 patentes associadas (10400 patentes ao todo). De modo geral, aplicou-se um pré-processamento básico, com a remoção de *stop words* e pontuações.

Para este estudo decidiu-se pelo conjunto de dados com 16 avaliadores, visto que este possui mais patentes. Após o pré-processamento, efetuou-se a divisão do conjunto de dados para as etapas de treinamento e teste. No primeiro caso foram separadas 600 patentes por examinador e, no segundo caso, 200 patentes por examinador, totalizando 9.600 patentes no conjunto de treino e 3.600 patentes no conjunto de teste.

### 5.2 IMPLEMENTAÇÃO DO MÉTODO PROPOSTO

O método proposto neste trabalho foi desenvolvido utilizando a linguagem de programação Python<sup>®</sup> uma vez que esta possui bibliotecas disponíveis que facilitam diferentes atividades de implementação no cenário de Ciência de Dados, entre elas, o pré-processamento e a geração de *embeddings*. O método é composto de etapas, que interconectadas, possibilitam a recomendação de examinadores para a análise de patentes. De modo geral, ao inserir uma patente como documento de entrada, tem-se uma lista de sugestões com os examinadores identificados como os mais aptos para analisar uma patente de interesse.

O fluxo das etapas do método inicia com o pré-processamento e padronização de dados, definindo-se quais campos serão utilizados no aprendizado do modelo de recomendação. Na sequência, ocorre a divisão do conjunto de dados em treinamento e teste. Isto é realizado utilizando a função *train\_test\_split* disponível no módulo *model\_selection* da biblioteca *Scikit-Learn*<sup>®</sup>.

Em seguida, cada documento pertencente aos conjuntos de treinamento e teste devem ser representados como uma lista de *tokens* juntamente com a identificação desses tokens (*tags*). Para tal, a função *TaggedDocument* da biblioteca *gensim*<sup>®</sup> é utilizada. Os vetores resultantes representam os documentos em formato adequado para serem utilizados como entrada para o *framework Doc2vec*<sup>®</sup>.

Após isso, o vocabulário para o conjunto de dados de treino é construído utilizando a função *build\_vocab*, também da biblioteca *gensim*<sup>®</sup>, tendo como entrada para esta função o conjunto de documentos de treino por meio da utilização da função *TaggedDocument*.

Com o vocabulário construído e os documentos formatados adequadamente pode-se então realizar a etapa de treinamento. Isto ocorre através da função *train* do *framework Doc2vec*<sup>®</sup>. O *Doc2Vec*<sup>®</sup> tem como objetivo gerar o modelo semântico que será descrito por vetores densos (*embeddings*) capazes de responder a determinada demanda, ou seja, a busca por patentes similares.

Após a geração do modelo, segue a etapa de provimento de uma lista ordenada (*ranking*) de examinadores a partir de uma patente de interesse. Nesse sentido, a patente de entrada é tokenizada utilizando a função *word\_tokenize* da biblioteca *nlk*<sup>®</sup> e, em seguida, é efetuado o processo de inferência do vetor de entrada utilizando a função *infer\_vector* da biblioteca *gensim*<sup>®</sup>. O vetor inferido é, em seguida, comparado aos vetores das patentes do modelo de recomendação gerado na Etapa 3, utilizando a função *most\_similar* e, por fim, o modelo retorna uma lista ordenada, do mais similar para o menos similar, com os documentos (patentes) com seus respectivos examinadores e as suas respectivas similaridades.

Como passo final, utilizando a lista de patentes com os respectivos examinadores, ocorre a agregação ao nível de examinador. Ou seja, ocorre a contagem de quantas patentes cada examinador está associado e, então, é produzida a relação final ordenada (na forma de um *ranking*) com a sugestão de examinadores aptos à análise de determinada patente.

### 5.3 AVALIAÇÃO DO MÉTODO PROPOSTO

Para a avaliação do método proposto, foram estabelecidos alguns cenários de modo a verificar, a partir de algumas patentes do conjunto de teste, a aderência entre o examinador da patente de entrada e a lista agregada de examinadores produzida pelo modelo de recomendação.

A aderência a lista agregada de examinadores sugeridos em função de uma patente de entrada é definida da seguinte forma: i) são definidas as  $n$  patentes que devem ser retornadas como similares à patente de entrada; e ii) com o resultado das  $n$  patentes retornadas como sendo as mais similares à patente de entrada, é gerada uma lista agregada e ordenada de examinadores e suas respectivas quantidades de patentes avaliadas, do examinador com mais patentes avaliadas para o que possui menos patentes avaliadas.

Já a avaliação do resultado da lista agregada de examinadores se utiliza de uma variável  $k$  indicando que, se o examinador que avaliou a patente de entrada estiver dentro deste limite definido, então é dito que a resposta da recomendação foi bem sucedida. Caso contrário, não houve êxito. Neste trabalho, os valores de  $n$  foram estabelecidos em 25, 50, 75 e 100 e, para um teste inicial do desempenho das consultas, o valor de  $k$  foi definido como 3. Visto que o número de examinadores é relativamente pequeno, 16 ao todo, valores de  $k$  maiores do que 3 tendem a produzir uma resposta positiva que não seria obtida em um cenário real.

Para este primeiro teste foi utilizada como patente de entrada a que consta no Quadro 5. Também definiu-se  $k=3$  permitindo avaliar se o examinador da patente consta até a terceira posição da lista agregada de examinadores (Quadro 6) obtida a partir de  $n=25$ . Como se pode verificar no Quadro 6, o “Examinador 01” não consta na lista retornada que possui ao todo 8 (oito) examinadores. O teste foi então repetido para os demais valores de  $n$  do estudo, 50, 75 e 100. Os resultados estão integrados no Quadro 6. Em nenhum dos casos o “Examinador 01” ficou posicionado entre os primeiros  $k=3$  examinadores da lista, sendo a melhor posição para  $n=50$ , na oitava colocação.

Quadro 5 – Patente de entrada avaliada pelo Examinador 01 para o primeiro cenário

Examinador	Texto da Patente
Examinador 01	<i>"plugged honeycomb structure. a honeycomb structure body is constituted of a circumferential region including the outermost circumference in a cross section of the honeycomb structure body which is perpendicular to an extending direction of cells, and a central region excluding the circumferential region. the circumferential region includes a specific circumferential region in which pressure loss with soot when an amount of the deposited soot is 4 g/l is higher than the pressure loss with soot of the central region as much as 15% or more and in which an open frontal area of the circumferential region is the same as or larger than an open frontal area of the central region. in the cross section of the honeycomb structure body, a ratio of an area of the specific circumferential region is 5% or more to a total area of the circumferential region and the central region."</i>

Fonte: Elaborado pelo autor (2022).

Quadro 6 – Lista de agregação de examinadores com as respectivas quantidades de patentes avaliadas para  $n$  igual a 25 para o primeiro cenário

Examinador	Quantidade de patentes avaliadas
Examiner 13	14
Examiner 10	3
Examiner 07	2
Examiner 08	2
Examiner 02	1
Examiner 09	1
Examiner 14	1
Examiner 15	1
<b>Total Geral</b>	<b>25</b>

Fonte: Elaborada pelo autor (2022).

Quadro 7 - Posição de retorno na lista de recomendações ( $k$ ) para os diferentes valores de  $n$  testados para o primeiro cenário

n	25	50	75	100
Posição de retorno considerando $k=3$	Não consta	8	10	9

Fonte: Elaborado pelo autor (2022).

Seguindo o mesmo processo do teste do primeiro cenário e considerando os mesmos valores de  $n$  e  $k$ , apresenta-se o documento para este segundo cenário com a patente (examinador e texto) que consta no Quadro 8. Para  $n=25$  e levando-se em conta a recomendação do Quadro 9, o “Examinador 06” fica posicionado na primeira colocação tendo avaliado 8 patentes, estando assim dentro do intervalo de  $k=3$  definido previamente. Para os demais valores de  $n$ , 50, 75 e 100, pode-se ver que o “Examinador 06” se mantém no topo em todas as listas agregadas de recomendação (Quadro 10).

Quadro 8 – Patente de entrada avaliada pelo Examinador 06 para o segundo cenário

<b>Examinador</b>	<b>Patente</b>
Examinador 06	<i>"design apparatus and a method for generating an implementable description of a digital system. the present invention is a design apparatus compiled on a computer environment for generating from a behavioral description of a system comprising at least one digital system part, an implementable description for said system, said behavioral description being represented on said computer environment as a first set of objects with a first set of relations therebetween, said implementable description being represented on said computer environment as a second set of objects with a second set of relations therebetween, said first and second set of objects being part of a design environment."</i>

Fonte: Elaborada pelo autor (2022).

Quadro 9 - Lista agregada de examinadores com as respectivas quantidade de patentes avaliadas para  $n$  igual a 25 para o segundo cenário

<b>Examinador</b>	<b>Quantidade de patentes avaliadas</b>
Examiner 06	8
Examiner 13	7
Examiner 03	4
Examiner 04	3
Examiner 05	1
Examiner 14	1
Examiner 16	1
<b>Total Geral</b>	<b>25</b>

Fonte: Elaborada pelo autor (2022).

Quadro 10: Posição de retorno na lista de recomendações ( $k$ ) para os diferentes valores de  $n$  testados para o segundo cenário

<b><math>n</math></b>	<b>25</b>	<b>50</b>	<b>75</b>	<b>100</b>
Posição de retorno considerando $k=3$	1	1	1	1

Fonte: Elaborado pelo autor (2022).

Para o terceiro cenário e considerando os mesmos valores de  $n$  e  $k$ , segue a apresentação da patente (examinador e texto) que consta no Quadro 11. Para  $n=25$  e com base na recomendação do Quadro 12, o “Examinador 07” fica posicionado na segunda colocação com 5 patentes avaliadas, estando dentro do intervalo de  $k=3$ . Para os demais valores de  $n$ , 50, 75 e 100, pode-se verificar que o “Examinador 07” ocupa a primeira colocação (Quadro 13). Neste caso, percebe-se que o aumento de  $n$  impacta positivamente no resultado.

Quadro 11 – Patente de entrada avaliada pelo Examinador 07 para o terceiro cenário

Examinador	Patente
Examinador 07	<i>"condenser cylinder adapted for use in an air-conditioning circuit, more specifically the air-conditioning circuit of an automobile. the invention relates to a condenser receiver adapted to receive and house a fluid used in a cold loop of an air-conditioning circuit. the condenser receiver comprises an outer wall and an inner wall which define an intermediate space, said inner wall defining an inner space for housing the fluid. the intermediate space comprises a static component adapted to store and release a given quantity of heat in order to allow a heat exchange between the static component and the fluid contained in the inner space."</i>

Fonte: Elaborado pelo autor (2022).

Quadro 12: Lista agregada de examinadores com as respectivas quantidade de patentes avaliadas para  $n$  igual a 25 para o terceiro cenário

Examinador	Quantidade de patentes avaliadas
Examiner 01	6
Examiner 07	5
Examiner 12	5
Examiner 14	3
Examiner 04	2
Examiner 04	1
Examiner 06	1
Examiner 08	1
Examiner 09	1
Examiner 10	1
Examiner 11	1
Examiner 15	1
<b>Total Geral</b>	<b>25</b>

Fonte: Elaborada pelo autor (2022).

Quadro 13: Posição de retorno na lista de recomendações ( $k$ ) para os diferentes valores de  $n$  testados para o terceiro cenário

$n$	25	50	75	100
Posição de retorno considerando $k=3$	2	1	1	1

Fonte: Elaborado pelo autor (2022).

Continuando, o quarto cenário baseia-se na patente (examinador e texto) que consta no Quadro 14, utilizando para tal, os mesmos valores de  $n$  e  $k$ . Para  $n=25$  e considerando a recomendação do Quadro 15, o “Examinador 09” fica posicionado na primeira colocação com 16 patentes avaliadas, estando dentro do intervalo de  $k=3$ . Para os demais valores de  $n$ , 50, 75 e 100, pode-se verificar que o “Examinador 09” se mantém na primeira colocação (Quadro 16).

Quadro 14 – Patente de entrada avaliada pelo Examinador 09 para o quarto cenário

Examinador	Patente
Examinador 09	<i>"method and apparatus of forming a coating film on an inner surface of a panel of a cathode ray tube. the present invention is directed to a method of and apparatus for forming a coating film on an inner surface of a panel of a cathode ray tube. in this case, slurries are injected into the panel, the panel is rotated on its own axis so as to uniformly coat the slurries on the panel and a recovering hood is moved in an opposing relation to the panel in synchronism with the revolution of the panel to thereby recover the slurries. therefore, particularly when the slurries are recovered, the slurries can be prevented from being scattered, thereby avoiding the surroundings from being splattered."</i>

Fonte: Elaborado pelo autor (2022).

Quadro 15: Lista agregada de examinadores com as respectivas quantidade de patentes avaliadas para  $n$  igual a 25 para o quarto cenário

Examinador	Quantidade de patentes avaliadas
Examiner 09	16
Examiner 05	2
Examiner 04	1
Examiner 06	1
Examiner 07	1
Examiner 11	1
Examiner 14	1
Examiner 15	1
Examiner 16	1
<b>Total Geral</b>	<b>25</b>

Fonte: Elaborada pelo autor (2022).

Quadro 16: Posição de retorno na lista de recomendações ( $k$ ) para os diferentes valores de  $n$  testados para o quarto cenário

$n$	25	50	75	100
Posição de retorno ( $k$ )	1	1	1	1

Fonte: Elaborado pelo autor (2022).

Por fim, para o quinto e último cenário e considerando os mesmos valores de  $n$  e  $k$ , segue a apresentação da patente (examinador e texto) que consta no Quadro 17. Para  $n=25$  e levando-se em conta as recomendações do Quadro 18, o “Examinador 12” fica posicionado na quarta colocação com apenas 1 patente avaliada estando fora do intervalo de  $k=3$ . Para os demais valores de  $n$ , 50, 75 e 100, verifica-se que o “Examinador 12” sobe 1 posição passando à terceira colocação em todas as situações (Quadro 19). Novamente, percebe-se um incremento do resultado para  $n=50$ , todavia, não produzindo uma melhora para além deste valor.

Quadro 17 – Patente de entrada avaliada pelo Examinador 12 para o quinto cenário

Examinador	Patente
Examinador 12	<i>"texture material for covering a repaired portion of a textured surface. a texture material formulated substantially to match a pre-existing texture pattern. the texture material comprises a carrier, a binder, a filler, and a ph change material. the texture material changes from a wet form to a solidified form when exposed to air. a ph of the texture material changes as the texture material changes from the wet form to the solidified form. an appearance of at least one appearance characteristic of the at least one material forming the ph change material changes as the ph of the texture material changes. an appearance characteristic of the texture material associated with the at least one material forming the ph change material in at least one of the wet form and the solidified form differs from a similar appearance characteristic of the pre-existing texture pattern."</i>

Fonte: Elaborada pelo autor (2022).

Quadro 18: Lista agregada de examinadores com as respectivas quantidade de patentes avaliadas para  $n$  igual a 25 - quinto cenário

Examinador	Quantidade de patentes avaliadas
Examiner 10	7
Examiner 14	3
Examiner 04	3
Examiner 12	1
Examiner 01	1
Examiner 03	1
Examiner 05	1
Examiner 06	1
Examiner 09	1
<b>Total Geral</b>	<b>25</b>

Fonte: Elaborada pelo autor (2022).

Quadro 19: Posição de retorno na lista de recomendações ( $k$ ) para os diferentes valores de  $n$  testados para o quinto cenário

$n$	25	50	75	100
Posição de retorno ( $k$ )	4	3	3	3

Fonte: Elaborado pelo autor (2022).

Nesta seção, foram realizados cinco testes, considerando para a variável  $n$  os valores de 25, 50, 75 e 100 patentes de retorno para determinada consulta, ou seja, determina quantas respostas serão providas para uma patente de entrada obtida a partir do conjunto de teste. O objetivo final dos testes consiste na verificação se o método proposto de recomendação utilizando vetores densos através do *Doc2Vec*<sup>®</sup> é capaz de recomendar, dentro de um intervalo  $k$ , examinadores aptos para a avaliação de uma determinada patente. Para tal, a variável  $k$  foi definida em 3, conforme justificativa já apresentada.

Analisando de modo geral os cenários, percebe-se que nos testes realizados para o primeiro cenário, o examinador da patente de entrada não figura entre as  $k$  primeiras sugestões da lista agregada. A melhor posição do “Examinador 01” foi a oitava colocação para  $n=50$ , não sendo adequado quanto a resposta esperada. Uma investigação mais detalhada seria necessária, tentando identificar o perfil do “Examinador 01” considerando o conjunto de suas patentes utilizadas para o treinamento. Neste sentido, informações adicionais poderiam ser obtidas para verificar, por exemplo, se a patente de entrada atribuída ao referido examinador está relacionada com o seu perfil, ou representa uma avaliação esporádica em função de uma demanda em particular, ou seja, não aderente ao seu perfil.

Já para os testes do segundo e quarto cenários, independente do valor de  $n$ , os examinadores identificados pelos códigos 06 e 09 ocuparam as primeiras posições no *ranking* recomendado pelo método proposto. Conclui-se que nestas duas situações o  $n$  não impacta nos resultados que são positivos. No terceiro cenário, todas as recomendações realizadas pelo método incluem o “Examinador 07” dentro do limite de  $k=3$  definido. Todavia, nesta situação o aumento do valor de  $n$  para 50 promoveu um melhor resultado. Incrementos subsequentes, por outro lado, mantiveram o resultado de  $n=50$ .

Por fim, para os testes realizados no quinto cenário, o “Examinador 12” indicado na patente de entrada, aparece na quarta posição da lista de recomendações de possíveis avaliadores para  $n=25$ . Portanto, fora do ponto de corte de  $k=3$ . Já para os demais valores de  $n$ , 50, 75 e 100, o examinador aparece uma posição à frente, indicando novamente que um  $n$  com valor intermediário, neste caso 50, pode conduzir a recomendações mais assertivas.

## 6. CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

O contínuo avanço no desenvolvimento de novos algoritmos, técnicas, abordagens, modelos e novos métodos têm contribuído significativamente para o avanço na construção de sistemas de recomendação sofisticados capazes de melhorar os diferentes processos de recomendação a que são destinados. Várias abordagens têm sido utilizadas para as mais diversas finalidades, por exemplo, a recomendação de examinadores.

Neste trabalho, foi proposto e desenvolvido um método voltado à recomendação de examinadores para a análise de patentes. O método possibilita, a partir de uma patente de entrada, comparar o vetor denso (*embedding*) da mesma com os vetores densos (*embeddings*) que representam o modelo de recomendação. Como resposta, as patentes mais similares são retornadas permitindo que, após a análise dos examinadores vinculados às patentes, seja estabelecido um ranking sugerindo examinadores aptos à análise da patente que gerou a demanda.

Em termos da avaliação do método proposto, os cinco cenários apresentados na seção 5.3 resumem o desempenho do método. Em cada um dos cenários, o método proposto foi avaliado considerando as variáveis  $n$  e  $k$ . Os valores de  $n$  foram estipulados em 25, 50, 75 e 100 patentes de retorno para uma determinada consulta (patente de entrada) com  $k$  igual a 3. Ou seja, para cada  $n$ , objetiva-se identificar as patentes mais similares levando-se em conta uma entrada  $e$ , a partir disso, gerar uma lista ordenada de examinadores de modo que esta lista possa ser avaliada segundo um  $k$  definido.

De modo geral, os resultados apresentados nestes cenários são consistentes com o esperado do método de recomendação de examinadores no contexto de análise de patentes. Quanto a variável  $n$  verificou-se que valores intermediários, levando-se em conta os que foram testados, contribuem para a recomendação de um *ranking* de examinadores mais adequado. Isso pode ser verificado nos cenários três e cinco, visto que o aumento de  $n$ , melhorou a recomendação. Em relação a variável  $k$  os resultados também se apresentam adequados, porém, são inconclusivos visto que no conjunto de dados existem somente 16 examinadores. Em um cenário real este número seria muito mais elevado. Todavia, recomendar uma lista ordenada com  $k$  maior do que 10 ou 15 examinadores não parece ser muito produtivo, uma vez que irá demandar bastante trabalho do tomador de decisão que precisa vincular/atribuir determinada patente a um examinador.

As alternativas para o desenvolvimento de um método voltado a recomendação de examinadores para a análise de patentes são inúmeras e podem ser continuamente aperfeiçoadas. Neste sentido, algumas possibilidades de trabalhos futuros foram identificadas. Cita-se, por exemplo, a evolução do modelo de recomendação para utilização de redes neurais profundas baseadas no conceito de *transformers* para produzir vetores densos contextualizados. Outra evolução plausível diz respeito à evolução do *ranking* de examinadores. Neste trabalho, a similaridade das patentes de retorno não foi considerada, somente a quantidade de patentes previamente avaliadas por determinado examinador. Estudos neste sentido poderiam impactar positivamente na qualidade do *ranking* de examinadores.

Por fim, em um sistema real, seria importante sugerir ou recomendar o examinador da lista agregada que está menos sobrecarregado, isto é, aquele cuja quantidade de patentes avaliadas é a menor entre os examinadores presentes no *ranking*. Ou seja, poderia existir um módulo de balanceamento de carga evitando o acúmulo de trabalho para determinados examinadores.

## REFERÊNCIA

- ABBAS, Assad; ZHANG, Limin; KHAN, Samee U. A literature review on the state-of-the-art in patent analysis. **World Patent Information**, v. 37, p. 3-13, 2014.
- AGGARWAL, Charu C. et al. **Recommender systems**. Cham: Springer International Publishing, 2016.
- ALMEIDA, Felipe; XEXÉO, Geraldo. Word embeddings: A survey. **arXiv preprint arXiv:1901.09069**, 2019.
- BAI, Wenlei et al. Collaborative Filtering Auto-Encoders for Technical Patent Recommending. **IEICE Transactions on Information and Systems**, v. 104, n. 8, p. 1258-1265, 2021.
- BIAGIOLI, Mario. Patent republic: representing inventions, constructing rights and authors. **Social Research: An International Quarterly**, v. 73, n. 4, p. 1129-1172, 2006.
- BLIND, Knut; POHLISCH, Jakob; ZI, Aikaterini. Publishing, patenting, and standardization: Motives and barriers of scientists. **Research Policy**, v. 47, n. 7, p. 1185-1197, 2018.
- BURKHARDT, Dominik; DILEXIT, Erik. David and Goliath: Patent litigation and its challenges for SMEs. 2021.
- CHA, Keum Gang et al. Performance of Recommender Systems: Based on Content Navigator and Collaborative Filtering. **arXiv preprint arXiv:1909.08219**, 2019.
- CHEN, Ming et al. Using document embedding techniques for similar bug reports recommendation. In: **2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)**. IEEE, 2018. p. 811-814.
- CHEN, Ying et al. SIMPLE: a strategic information mining platform for licensing and execution. In: **2009 IEEE International Conference on Data Mining Workshops**. IEEE, 2009. p. 270-275.
- DUMAIS, Susan T. Latent semantic analysis. **Annual Review of Information Science and Technology (ARIST)**, v. 38, p. 189-230, 2004.
- FU, Tao-yang; LEI, Zhen; LEE, Wang-Chien. Patent citation recommendation for examiners. In: **2015 IEEE international conference on data mining**. IEEE, 2015. p. 751-756.
- GARTEN, Justin et al. Combining distributed vector representations for words. In: **Proceedings of the 1st workshop on vector space modeling for natural language processing**. 2015. p. 95-101.
- GORSHKOV, Artur. Patenting in Biotechnology. **Available at SSRN 3775173**, 2021.
- HALL, Bronwyn H. Patents and patent policy. **Oxford Review of Economic Policy**, v. 23, n. 4, p. 568-587, 2007.
- HAWASHIN, Bilal et al. Novel weighted interest similarity measurement for

recommender systems using rating timestamp. In: **2019 Sixth International Conference on Software Defined Systems (SDS)**. IEEE, 2019. p. 166-170.

HORSTMANN, Ignatius; MACDONALD, Glenn M.; SLIVINSKI, Alan. Patents as information transfer mechanisms: To patent or (maybe) not to patent. **Journal of Political Economy**, v. 93, n. 5, p. 837-858, 1985.

JAIN, Sonal; KHANGAROT, Harshita; SINGH, Shivank. Journal recommendation system using content-based filtering. In: **Recent developments in machine learning and data analytics**. Springer, Singapore, 2019. p. 99-108.

KALYAN, Katikapalli Subramanyam; SANGEETHA, Sivanesan. SECNLP: A survey of embeddings in clinical natural language processing. **Journal of biomedical informatics**, v. 101, p. 103323, 2020.

KOCHER, Mirco; SAVOY, Jacques. Distance measures in author profiling. **Information processing & management**, v. 53, n. 5, p. 1103-1119, 2017.

KRESTEL, Ralf et al. A survey on deep learning for patent analysis. **World Patent Information**, v. 65, p. 102035, 2021.

LATTMAN, E. E. Patent History. The Editor's Corner. Wiley Online Library, n. 2, 1996.

LE, Quoc; MIKOLOV, Tomas. Distributed representations of sentences and documents. In: **International conference on machine learning**. PMLR, 2014. p. 1188-1196.

LU, Jie et al. Recommender system application developments: a survey. **Decision Support Systems**, v. 74, p. 12-32, 2015.

CECHETT, Fernando. Modelagem de programação dinâmica para sequenciamento de planos curriculares no ensino superior. 2021.

MAO, Guozhu et al. Technology status and trends of industrial wastewater treatment: A patent analysis. **Chemosphere**, v. 288, p. 132483, 2022.

PARK, Kwangil; HONG, June Seok; KIM, Wooju. A methodology combining cosine similarity with classifier for text classification. **Applied Artificial Intelligence**, v. 34, n. 5, p. 396-411, 2020.

SHALABY, Walid; ZADROZNY, Wlodek. Patent retrieval: a literature review. **Knowledge and Information Systems**, v. 61, n. 2, p. 631-660, 2019.

SILVA, Thales do Nascimento da. **Um Modelo de Recomendação de Colaboradores Voltado à Execução de Tarefas no Cenário de Crowdsourcing**. 2022. Qualificação de Doutorado (Doutorado) - Programa de Pós-Graduação de Engenharia e Gestão do Conhecimento, Universidade Federal de Santa Catarina, Florianópolis, 2022.

SJÖGREN, Rickard et al. Multivariate patent analysis—Using chemometrics to analyze collections of chemical and pharmaceutical patents. **Journal of Chemometrics**, v. 34, n. 1, p. e3041, 2020.

TANG, Jie et al. Patentminer: topic-driven patent analysis and mining. In: **Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and**

**data mining**, 2012. p. 1366-1374.

TASNIM, Zerín et al. Political Ideology Prediction from Bengali Text Using Word Embedding Models. In: **2021 International Conference on Emerging Smart Computing and Informatics (ESCI)**. IEEE, 2021. p. 724-727.

TRAPPEY, Amy JC et al. Intelligent patent recommendation system for innovative design collaboration. **Journal of Network and Computer Applications**, v. 36, n. 6, p. 1441-1450, 2013.

TRAPPEY, Amy JC et al. Patent value analysis using deep learning models—The case of IoT technology mining for the manufacturing industry. **IEEE Transactions on Engineering Management**, v. 68, n. 5, p. 1334-1346, 2019.

TRAPPEY, Amy; TRAPPEY, Charles V.; HSIEH, Alex. An intelligent patent recommender adopting machine learning approach for natural language processing: A case study for smart machinery technology mining. **Technological Forecasting and Social Change**, v. 164, p. 120511, 2021.

TRIPPE, Anthony J. Patinformatics: Tasks to tools. **World Patent Information**, v. 25, n. 3, p. 211-221, 2003.

ULČAR, Matej; ROBNIK-ŠIKONJA, Marko. High quality ELMo embeddings for seven less-resourced languages. **arXiv preprint arXiv:1911.10049**, 2019.

YANG, Yun Yun et al. Enhancing patent landscape analysis with visualization output. **World Patent Information**, v. 32, n. 3, p. 203-220, 2010.

YUE, Wang; LI, Lei. Sentiment Analysis using Word2vec-CNN-BiLSTM Classification. In: **2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)**. IEEE, 2020. p. 1-5.

YUE, Wang; LI, Lei. Sentiment Analysis using Word2vec-CNN-BiLSTM Classification. In: **2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)**. IEEE, 2020. p. 1-5.

ZHANG, Yanping et al. Semantic Based Heterogeneous Information Network Embedding for Patent Citation Recommendation. In: **2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE)**. IEEE, 2020. p. 518-527.