CONFERENCIA INTERNACIONAL BIREDIAL-ISTEC 25
08-10 de octubre de 2025 Brasilia • BRASIL

IMPLEMENTAÇÃO DE UMA ESTRATÉGIA DE BUSCA SEMÂNTICA EM CONJUNTOS DE DADOS BIBLIOGRÁFICOS DE PATENTES

Raulivan Rodrigo da Silva
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Brasil | raulivan@cefetmg.br
https://orcid.org/XXXX-XXXX-XXXX-XXXX

Thiago Magela Rodrigues Dias
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG), Brasil | thiagomagela@cefetmg.br
https://orcid.org/XXXX-XXXX-XXXX-XXXX

Washington Luís Ribeiro de Carvalho Segundo
Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict), Brasil | washingtonsegundo@ibict.br
https://orcid.org/XXXX-XXXX-XXXX-XXXX

DOI: *10.22477/xiv.biredial.402*

EJE TEMÁICO: *Infraestrutura tecnológica*

RESUMEN

Este estudo propõe uma estratégia para a realização de busca semântica para a recuperação de dados bibliográficos de patentes, visando propor uma alternativa aos métodos tradicionais de recuperação da informação, como a pesquisa booleana. Para alcançar o objetivo, foi adotada uma abordagem metodológica qualitativa e exploratória, com ênfase no desenvolvimento e avaliação de estratégias computacionais eficientes para aplicar a busca semântica. Como contribuição mais relevante, o presente trabalho apresenta o desenvolvimento de um framework implementado em Python, capaz de aplicar técnicas de busca semântica em conjuntos bibliográficos de patentes armazenados localmente. Essa solução constitui uma alternativa relevante aos modelos tradicionais. Os resultados indicam que a busca semântica otimiza o processo de recuperação de documentos de patentes. Assim, o estudo contribui para o avanço de estratégias computacionais voltadas à recuperação da informação em documentos de patentes.

Palabras-clave: Patentes; Busca; Semântica; Espacenet.

ABSTRACT

(ver resumo em inglês com autores)

Keywords: Patents; Search; Semantics; Espacenet.

INTRODUÇÃO

Desde a formalização da propriedade intelectual por meio dos documentos de patentes, é perceptível um crescimento contínuo no número de depósitos ao longo dos anos. O avanço acelerado das tecnologias e o surgimento constante de novos dispositivos, aplicações e meios digitais têm impulsionado a criação de versões aprimoradas de funcionalidades já existentes, bem como o desenvolvimento de soluções inovadoras.

De acordo com dados da Organização Mundial da Propriedade Intelectual (OMPI), foram registrados, em âmbito global, 3,4 milhões de pedidos de patentes em 2021, representando um crescimento de 3,6% em relação aos anos anteriores (World Intellectual Property Organization [WIPO], 2022). No contexto nacional, conforme informações do Instituto Nacional da Propriedade Industrial (INPI), o Brasil contabilizou 20.343 depósitos de patentes entre janeiro e setembro de 2024, refletindo um aumento de 1,5% em comparação ao mesmo período de 2023 (Instituto Nacional da Propriedade Industrial [INPI], 2024). Tais dados evidenciam o crescente interesse das organizações em proteger suas inovações tecnológicas, com vistas à consolidação de vantagens competitivas e à fidelização de seus mercados consumidores (Amadei & Torkomian, 2009).

Neste cenário, o campo da Propriedade Intelectual assume um papel estratégico na promoção da inovação e no fomento ao desenvolvimento tecnológico em diferentes áreas do conhecimento. Em particular, os documentos de patentes destacam-se como fontes ricas de informação sobre o estado da técnica, tendências de mercado e estratégias de atuação de empresas e instituições de pesquisa. A estrutura desses documentos exige uma descrição técnica detalhada da invenção, incluindo desenhos, esquemas, e reivindicações que definem o escopo legal da tecnologia protegida (Sanz-Casado, 2006; Nascimento & Spezialili, 2020; Rezende et al., 2023). Conforme argumentam Reymond e Quoniam (2018), os documentos de patentes concentram uma quantidade significativamente superior de informações técnicas em comparação com artigos científicos, o que reforça seu valor informacional no contexto da pesquisa científica e tecnológica.

Apesar de sua importância, a recuperação de documentos de patentes ainda enfrenta desafios consideráveis, especialmente no que diz respeito à manipulação de grandes volumes de dados, às restrições operacionais de acesso, e às limitações impostas pelos sistemas de busca disponíveis nos repositórios. Em conformidade com o princípio da publicidade dos pedidos de patentes, tais documentos são disponibilizados em repositórios de acesso público disponíveis na internet, mantidos por escritórios de propriedade industrial. No Brasil, esse papel é desempenhado pelo repositório do INPI (Lei n. 9.279, de 14 de maio de 1996, 1996). Em nível internacional, destacam-se plataformas como a Espacenet, que consolida dados de múltiplos escritórios de patentes, bem como serviços independentes como o Google Patents (Brandão, 2016).

Esses repositórios, embora amplamente acessíveis, diferem em diversos aspectos, dentre eles à arquitetura de seus sistemas de busca. De modo geral, é possível encontrar nos repositórios de patentes dois tipos principais: busca simples e busca avançada. A busca simples oferece campos pré-configurados nos quais os usuários inserem termos específicos, limitando, em certa medida, a flexibilidade das consultas. Por sua vez, a busca avançada requer conhecimento técnico por parte do usuário, sendo baseada na formulação de expressões lógicas e uso de operadores booleanos, como "AND" e "OR", ou até mesmo linguagens específicas como a *Contextual Query Language* (CQL).

Diante desse panorama, este trabalho tem como objetivo propor uma estratégia de busca aprimorada, orientada à semântica dos termos utilizados nas consultas. A proposta tem como objetivo aumentar a precisão e a relevância dos resultados por meio da interpretação do significado contextual de palavras e expressões, indo além da correspondência exata de palavras-chave. Ao considerar as relações semânticas e o contexto informacional da pesquisa para oferecer resultados mais relevantes para o pesquisador, contribuindo para o avanço das práticas de recuperação da informação no contexto da Propriedade Intelectual.

FUNDAMENTAÇÃO TEÓRICA

A recuperação eficaz de documentos de patentes em um repositório de dados requer a formulação de uma estratégia de busca bem estruturada. De acordo com Lopes (2002), uma estratégia de busca consiste no estabelecimento de um conjunto de regras que viabilizam a obtenção de respostas, em uma base de dados, para uma pergunta previamente formulada. Assim, uma estratégia de busca tem por objetivo definir os termos, operadores e atributos apropriados para localizar informações relevantes em um banco de dados.

Os repositórios online de patentes oferecem interfaces que possibilitam aos pesquisadores executarem essas estratégias de busca viabilizando a recuperação de documentos de patentes. Nesta seção, são conceituados os principais tipos de pesquisa disponíveis nesses repositórios.

O método mais simples e amplamente difundido é a busca por palavras-chave, que consiste na identificação de palavras-chave em atributos específicos da patente, tais como título, resumo ou inventor. Cada repositório de patente define quais os campos disponíveis para consulta, os critérios de correspondência (por exemplo: igualdade exata ou presença do termo) e fornecem caixas de entrada para inserção dos termos desejados (Lopes, 2002; Pires et al., 2020). Este método de pesquisa é mais simples e recomendado para pesquisadores menos experientes. Um exemplo dessa abordagem seria: Campo: "Título da Patente"; Critério de comparação: "="; Palavra-chave: "processo e dispositivo para testar a presença de microrganismos".

Outra abordagem bastante utilizada é a pesquisa booleana, também denominada pesquisa por expressão. Esse tipo de busca, embora mais complexa, proporciona maior flexibi-

lidade e precisão na formulação das consultas. Nessa modalidade, o pesquisador emprega operadores booleanos — geralmente os clássicos <AND>, <OR>, <NOT> ou <ANDNOT> — para combinar ou excluir termos, conforme a lógica de interesse (Lopes, 2002; Manglano & Zulueta, 2007). Por exemplo:((TITULO = “processo e dispositivo para testar a presença de microrganismos”) OR (TITULO = “sistema para testar a presença de microrganismos”)) AND (DEPOSITANTE= “Centro Federal de Educação Tecnológica de Minas Gerais”).

Além desses métodos de pesquisa, alguns repositórios oferecem funcionalidades mais especializadas, que, embora não amplamente adotadas, ampliam o escopo das buscas. Dentre essas funcionalidades, é possível destacar a pesquisa por citações de patentes, que permite identificar patentes que citam outras patentes. Outra funcionalidade específica é a busca por compostos químicos, que pode ser realizada por meio de fórmulas moleculares ou desenhos estruturais da substância desejada (Manglano & Zulueta, 2007).

Com o avanço computacional e o desenvolvimento de algoritmos de aprendizado, surgem os mecanismos de busca semântica em bases de dados. A busca semântica textual tem como objetivo recuperar informações a partir do significado contextual das expressões utilizadas, considerando as relações semânticas entre termos e a estrutura sintática dos enunciados (Liu, 2019).

Por exemplo, ao realizar uma busca com a palavra “condutor”, o sistema é capaz de interpretar, com base no contexto, se o termo se refere a um motorista, a uma pessoa que conduz algo, ou à propriedade de materiais condutores de eletricidade com base na análise dos demais termos. Em síntese, a semântica, no contexto da linguagem natural, dedica-se ao estudo do significado e da interpretação de palavras, frases ou expressões, considerando o contexto em que estão inseridas.

No contexto computacional, diversos algoritmos de Deep Learning aplicados ao Processamento de Linguagem Natural (PLN) vêm sendo empregados para viabilizar a implementação de buscas semântica de dados. Dentre os modelos mais relevantes, destaca-se o BERT (*Bidirectional Encoder Representations from Transformers*), desenvolvido pelo Google. Este modelo é capaz de compreender o contexto de uma consulta textual de forma bidirecional, ou seja, levando em consideração tanto os termos que antecedem quanto aqueles que sucedem uma determinada palavra na sentença (Devlin et al., 2018).

O BERT foi projetado para o pré-treinamento de representações profundas e bidirecionais a partir de grandes volumes de textos não rotulados, sendo capaz de codificar contextos complexos ao concatenar diferentes partes da entrada em uma única sequência. Esta sequência é iniciada com o token especial [CLS] e separada com o token [SEP], permitindo a fusão de três vetores de incorporação: de palavras, de posições e de segmentos. Tal arquitetura resulta em maior precisão e relevância nos resultados obtidos (Barros, 2022).

Outro aspecto de destaque é o fato de que o BERT foi treinado com conjuntos extensos de textos em diversos idiomas, incluindo o português brasileiro, o que amplia sua capacidade de interpretação e torna sua aplicação eficaz em sistemas de busca semântica. Voltados à análise de linguagem natural em contextos diversos, inclusive em bases de dados técnico-científicas e documentos de patentes.

METODOLOGIA

O presente estudo possui natureza de pesquisa aplicada, desenvolvida com orientação epistêmica hipotético dedutiva alinhada com o problema e objetivo anteriormente apresentados. O foco está em gerar conhecimento prático voltado para a resolução de problemas específicos no campo da propriedade industrial, particularmente recuperação semântica de dados bibliográficos de patentes.

Adicionalmente, a pesquisa adota uma orientação metodológica exploratória e descritiva, pois busca explorar estratégias e ferramentas para recuperar os dados.

Quanto aos procedimentos, a pesquisa combina métodos de pesquisa documental e bibliográfica, pois baseia-se na análise de documentos de patentes disponíveis publicamente, além da revisão da literatura pertinente à repositórios de patentes.

DESENVOLVIMENTO

A estratégia de busca semântica proposta neste estudo tem como objetivo a recuperação de patentes em uma base de dados local, a partir de descrições formuladas em linguagem natural. Essa abordagem tem como objetivo superar as limitações das buscas tradicionais baseadas em palavras-chave, ao considerar o significado contextual dos termos empregados nas consultas. Para a construção da base de dados local, de acordo com abordagem metodológicas apresentadas por Silva et al. (2024), foram coletadas patentes depositadas no Brasil entre os anos de 1900 e 2024, disponibilizadas no repositório Espacenet. Esse conjunto de dados constituirá a base de dados local de patentes usada neste estudo.

Para a implementação da estratégia de busca, foi selecionada a linguagem de programação Python, amplamente adotada nas áreas de Ciência da Informação e Ciência da Computação, sobretudo por sua simplicidade sintática, vasto ecossistema de bibliotecas e alto desempenho em tarefas de mineração e análise de dados. Ademais, o Python se destaca, em 2025, como uma das linguagens mais utilizadas no setor tecnológico, o que reforça sua relevância para o desenvolvimento da ferramenta proposta.

A base computacional da busca semântica foi estruturada a partir do modelo BERT (*Bidirectional Encoder Representations from Transformers*), desenvolvido pela Google, classificado

como uma LLM *(Large Language Model*), ou seja, um modelo de aprendizado profundo treinado com grandes volumes de dados textuais. Em particular, foi adotada a variante S-BERT (*Sentence-BERT*), uma extensão do BERT otimizada para a geração de *embeddings* semânticos de sentenças. Essa implementação é viabilizada por meio da biblioteca Python SentenceTransformers (<https://pypi.org/project/sentence-transformers/>).

A biblioteca SentenceTransformers permite o cálculo de representações vetoriais (*embeddings*) de frases e textos, sendo especialmente eficaz na identificação de sinônimos e na aproximação semântica entre descrições textuais, o que a torna adequada para tarefas de busca em bases documentais complexas, como a de patentes. Além disso, o modelo oferece suporte a mais de 100 idiomas, incluindo o português brasileiro, o que garante maior aderência às particularidades linguísticas da base nacional de patentes. O repositório oficial da ferramenta S-BERT disponibiliza uma variedade de modelos pré-treinados, otimizados para diferentes tarefas de PLN, e pode ser acessado em: <https://www.sbert.net/index.html>.

O modelo pré-treinado selecionado para a implementação da estratégia de busca semântica foi o “distiluse-base-multilingual-cased-v1”, pertencente à família sentence-transformers. Este modelo mapeia sentenças e parágrafos para um espaço vetorial denso de 512 dimensões, sendo particularmente adequado para tarefas como agrupamento de documentos e recuperação semântica de informações (Reimers & Gurevych, 2019).

A implementação da busca semântica de patentes requer uma etapa inicial de preparação do banco de dados, denominada pré-processamento de dados, que visa estruturar as informações em um formato compatível para a análise semântica. Neste estudo, foi considerado apenas os campos título e resumo das patentes, uma vez que são os elementos que concentram as descrições mais relevantes do conteúdo técnico.

O pré-processamento é dividido em quatro etapas. A primeira etapa (i) consiste na seleção estruturada das patentes da base de dados local, extraindo os seguintes atributos: número da patente, título e resumo. Em seguida, realiza-se a etapa (ii) de limpeza de dados, na qual são excluídos todos os registros que não contenham os campos obrigatórios (número, título ou resumo), bem como aqueles que apresentem valores inconsistentes ou inválidos para os campos classificados como obrigatórios. A etapa subsequente (iii) corresponde à transformação dos dados, que envolve a conversão de todos os caracteres para minúsculas e a remoção de palavras irrelevantes (*stopwords*), a fim de reduzir ruídos semânticos. Finalmente, na etapa (iv) de redução dos dados, em que os campos “título” e “resumo” são concatenados em uma única descrição, resultando em uma base tabular contendo apenas dois atributos: número da patente e descrição consolidada.

Concluída a preparação do conjunto de dados, dá início a fase de processamento semântico, na qual o conjunto de dados tratado é convertido em vetores numéricos (*embeddings*) por meio do modelo distiluse-base-multilingual-cased-v1. Essa representação vetorial captura

o significado das sentenças e suas relações semânticas, viabilizando sua posterior análise por algoritmos de aprendizado de máquina. É importante destacar que, dada a natureza computacionalmente complexa do processo de geração de embeddings, especialmente em bases extensas, foi adotada uma estratégia de indexação por ano de depósito. Com isso, os dados são segmentados anualmente, reduzindo o volume de documentos por lote e otimizando o desempenho computacional.

Concluída as fases de pré-processamento e processamento, o conjunto de dados fica otimizado para execução de processamento semânticos como busca e agrupamentos. Para a execução da busca semântica, o texto usando para consulta, por exemplo, "sistema de descongelamento para aparelho de refrigeração", também passa por um pipeline de pré-processamento, incluindo a conversão para minúsculas, a remoção de *stopwords* e a transformação em *embedding*. Em seguida, o vetor resultante é comparado com os vetores das patentes indexadas por ano, utilizando medidas de similaridade semântica. O modelo SentenceTransformer retorna, para cada patente, um escore de relevância em relação ao termo consultado, variando entre 0.0 (baixa similaridade) e 1.0 (alta similaridade).

RESULTADOS

Como resultado deste estudo, foi desenvolvido um conjunto de algoritmos, implementados na linguagem de programação Python, capazes de realizar o pré-processamento, a indexação e a busca semântica em conjuntos de patentes de diferentes dimensões. A aplicação da estratégia de busca semântica demonstrou-se eficaz na recuperação de documentos relevantes a partir de descrições textuais formuladas em linguagem natural. Os resultados obtidos indicam que as respostas com índice de similaridade inferior a 30% (0.3) podem ser descartadas, por apresentarem baixo grau de relevância em relação ao termo de busca, considerados como ruído informacional.

O corpus utilizado neste estudo é composto por 831.706 registros de patentes, totalizando aproximadamente 1 gigabyte de dados textuais. O tempo necessário para a execução completa das fases de pré-processamento e geração dos vetores de *embeddings* foi de aproximadamente 204 horas. Este processamento resultou na criação de 51 índices distintos, segmentados por ano de depósito das patentes, sendo realizado em um ambiente computacional com 16 GB de memória RAM e processador Intel(R) Core(TM) i5-6300U CPU @ 2.40GHz. Destacando que esse processo é necessário ser executado apenas uma única vez, ou quando os dados considerados forem atualizados.

Por sua vez, a fase de busca semântica apresenta resultados satisfatórios. Embora envolva a varredura sequencial dos índices de *embeddings*, o algoritmo retorna resultados parciais em tempo real, proporcionando ao usuário a percepção de agilidade durante a consulta. Em termos de desempenho, a busca completa é concluída, em média, em menos de dois minutos,

evidenciando a viabilidade da solução proposta para ambientes de pesquisa que demandam respostas rápidas e contextualizadas.

A relevância dos resultados obtidos por meio da busca semântica também se mostrou satisfatória, uma vez que, em todos os casos analisados, as patentes retornadas apresentaram correspondência significativa com os termos expressos na consulta em linguagem natural. Essa coerência evidencia a eficácia do modelo na identificação de documentos de patentes.

A Tabela 1 apresenta uma amostragem representativa dos resultados gerados a partir da busca semântica com o seguinte enunciado: “Dispositivo ou sistema de refrigeração ou resfriamento”. A seleção contempla os registros com maiores índices de similaridade.

Tabela 1 – Amostragem de resultados da busca semântica

Relevância	Número	Descrição resumida
56,05	BR112015017785A2	1/1 resumo sistema de descongelamento para aparelho de refrigeração e unidade refrigerante um sistema de descongelamento inclui: um dispositivo de esfriamento que é disposto em um freezer, e inclui
55,34	BR0105644A	"EQUIPAMENTO DE REFRIGERAçãO APERFEIçOADO E SISTEMA DE REFRIGERAçãO APERFEIçOADO". Notadamente de um equipamento e respectivo sistema, através dos quais consegue-se vantagens e melhorias significati
55,08	BR9405086A	Sistema de refrigeração para aparelho de refrigeração
54,86	BR0303842B1	SISTEMA DE ABASTECIMENTO DE FORMAS DE GELO EM APARELHOS DE REFRIGERAÇÃO
54,48	BRPI0614107A2	MÉTODO E APARELHO PARA PROVER REFRIGERAÇçO A UM DISPOSITIVO. Um sistema para esfriar um ou mais dispositivos supercondutores discretos (21,22,23) e que um refrigerador primário (1) sub-esfria líquid
54,41	BR112015017791A2	sistema de descongelamento por sublimação e método de descongelamento por sublimação para aparelho de refrigeração
53,71	BRPI0601266A	"SISTEMA DE CONDENSAçãO APLICADO NO SISTEMA DE REFRIGERAçãO DE EQUIPAMENTOS DE REFRIGERAçãO E OU CONGELAMENTO". O presente privilégio revela um aparato que destina ser utilizado nos equipamentos de
52,85	BR102014023977A2	sistema conversível para congelamento e descongelamento de produtos alimentícios em aparelhos eletrodomésticos. a presente invenção pertence ao campo tecnológico dos aparelhos letrodomésticos, e ref
52,71	BR102014005557A2	Sistema de refrigeração de refrigerador apresentando um circuito de refrigeração secundário. A presente invenção refere-se a um sistema e método de refrigeração de refrigerador que proveem a refrige
52,31	BR9503236A	Sistema de degelo para aparelho de refrigeração
52,13	BR112015006534A2	1 / 1 resumo â refrigerador, e, mã todo para controlar o sistema de resfriamento de um refrigeradorâ se proporciona um refrigerador (2) compreendendo um sistema de resfriamento (4) sendo atravessado
51,96	BR112018004427B1	DISPOSITIVO E MÉTODO DE RESFRIAMENTO PARA O RESFRIAMENTO SECUNDÁRIO DE UM LINGOTE. A invenção refere-se a um dispositivo de resfriamento (7) e a um método de resfriamento para o resfriamento secundá
51,45	BR112015017789A2	1/1 resumo sistema de descongelamento para aparelho de refrigeração e unidade de resfriamento um sistema de descongelamento inclui: um dispositivo de resfriamento, que é disposto em um freezer, e in
51,24	BRPI0815928A2	equipamento para uso com um sistema que possui um componente de resfriamento criogênico, método utilizado com um sistema que possui um componente de resfriamento criogênico
51,10	BRPI0803841A2	SISTEMA PARA MOVIMENTAçãO DE UM CONJUNTO DE PRATELEIRAS DE EQUIPAMENTOS DE REFRIGERAçãO E EQUIPAMENTO DE REFRIGERAçãO. A presente invenção refere-se a um sistema capaz de prover um deslocamento em t
51,09	BR112022009060A2	MÁQUINA DE GELO DE RESFRIAMENTO DIRETO. Um aparelho de refrigeração inclui um compartimento de alimentos frescos para armazenar itens alimentícios em um ambiente refrigerado tendo uma temperatura al
50,84	BR112018013683A2	trata-se de um aparelho de refrigeração que inclui um compartimento de alimento fresco e um compartimento de congelador. uma máquina de produzir gelo com uma fôrma de gelo é disposta dentro do compa

A Tabela 1 apresenta, em sua primeira coluna, o percentual de similaridade entre cada patente recuperada e o termo de busca utilizado. A segunda coluna exibe o número de publicação da patente, enquanto a terceira coluna mostra, de forma truncada, os primeiros 200 caracteres da descrição da patente tal como informada no processo de depósito. Essa limitação visa otimizar a visualização dos dados apresentados na tabela.

É possível observar que todos os 17 registros exibidos demonstram relação semântica direta com o enunciado fornecido como critério de busca. O que reforça a eficácia da abordagem semântica na recuperação de documentos, uma vez que dispensa a formulação exata de palavras-chave, característica típica das abordagens tradicionais. O modelo utilizado é capaz de reconhecer variações dos termos, como refrigeração, refrigerado, refrigerante, mantendo a coerência do significado original, o que contribui para obter resultados mais relevantes para a busca.

De maneira complementar, a Tabela 2 apresenta os resultados classificados como de baixa relevância, ou seja, registros que, apesar de retornados pela busca semântica, apresentam baixa similaridade com o termo consultado, sendo considerados ruídos informacionais. Esses dados são essenciais para avaliar a precisão do modelo e delimitar o limite mínimo de similaridade aceitável para filtragem dos resultados em aplicações práticas.

Tabela 2 – Amostragem de resultados menos relevantes da busca semântica

Relevância	Número	Descrição resumida
27,88	BR8007413A	SISTEMA DE MANIPULACAO DE FILEIRA E DISPOSITIVO PRENDEDOR DE FILEIRA
27,74	BR8006433A	PROCESSO, APARELHO, SISTEMA E DISPOSITIVO PARA REFINACAO DE MATERIAIS FIBROSOS
27,19	BR7407402A	SISTEMA DE SUSPENSAO PARA UMA UNIDADE ATRELADA
27,12	BR7409009A	PROCESSO E DISPOSITIVO PARA TROCA DE MATERIAL ENTRE SISTEMAS HETEROGENEOS
26,34	BR7402474A	DISPOSITIVO E SISTEMA PARA DETETAR PARTICULAS SOLIDAS NUM FLUXO DE CORRENTE
26,27	BR7403478A	SISTEMA DE COLOCACAO DE APARELHO ECONOMIZADOR DE GASOLINA
26,00	BR7409067A	SISTEMA DE CONTROLE PARA EQUIPAMENTO DE MOLDAGEM DE SOLADOS PARA CALCADOS
25,70	BR7406594A	APERFEICOAMENTO EM APARELHO DE EVAPORACAO PARA USO NA CONCENTRACAO POR EVAPORACAO E/OU PURIFICACAO DE SOLUCOES E SISTEMA DE EVAPORACAO DE MULTIPLOS EFEITOS
25,66	BR7403450A	DISPOSITIVO E PROCESSO DE SUPRIMENTO DE FLUIDO A UM SISTEMA INFLAVEL PARA AMPARAR O OCUPANTE DE UM VEICULO
24,34	BR7405136A	PROCESSO E APARELHO PARA RESFRIAR UM MATERIAL QUENTE EM PARTICULAS
23,13	BR7402628A	SISTEMA DE FRITAR COM AGUA E AZEITE A TEMPERATURA VARIAVE
23,12	BR5401564U	PROCESSO DE INSTALACAO E ADAPTACAO DE FECHADURAS PARA GELADEIRAS
21,99	BR7403207A	MAQUINA ENSACADEIRA DE CARNES PARA OBTENCAO DE FRIOS EM GERAL
21,83	BR5401565U	PROCESSO DE INSTALACAO E ADAPTACAO DE FECHADURAS PARA GELADEIRAS
21,38	BR7403284A	DISPOSITIVO PARA A VENTILACAO E EXAUSTAO DE AMBIENTES EDIFICIOS E SIMILARES E DISPOSITIVO PARA A PERMUTA DE CALOR DE DOIS CIRCUITOS DE ELEMENTOS
21,31	BR7410986A	DISPOSITIVO DE BOMBA PARA CONTROLE DE PROPORCAO REGIME VELOCIDADE DA CIRCULACAO DE LIQUIDO NUM SISTEMA DE ENCANAMENTO
21,30	BR5403295U	CAIXA AUTO-ESTRUTURADA PARA REFRIGERADOR
21,13	BR7408722A	SISTEMA APERFEICOADO DE AUMENTO DE INTUICAO DE BOAS CONDICOES DE VOO SEM CONSULTA A APARELHO DE MEDIDA
21,03	BR7402890A	SISTEMA DE APLICACAO DE UNIDADES COM CARACTERISTICAS PROPRIAS DE PIEZOELETRICO EM FOGOES FOGAREIROS AQUECEDORES FORNOS E SEUS SIMILARES
20,57	BR5403359U	CAIXA AUTO-ESTRUTURADA PARA CONGELADORES
19,43	BR7506158A	SISTEMA DE PROTENSAO

Conforme evidenciado na Tabela 2, os resultados que apresentam percentual de similaridade inferior a 30% demonstram, de fato, baixa relevância em relação ao termo de busca utilizado. Esses registros incluem patentes que não abordam diretamente tecnologias ou dispositivos relacionados a sistemas de refrigeração ou resfriamento, sendo considerados ruído informacional no contexto da consulta semântica. Esses resultados reforçam a estratégia ado-

tada neste estudo, na qual apenas patentes com similaridade superior a 30% devem ser consideradas para fins analíticos, a fim de garantir maior precisão na recuperação da informação e relevância dos resultados obtidos.

CONCLUSÃO

A implementação da estratégia de busca semântica para recuperação de dados de patentes em base de dados locais apresentou resultados satisfatórios e tecnicamente viáveis, oferecendo uma alternativa relevante às abordagens tradicionais baseadas em palavras-chave. Bem como a estratégia de indexação por ano de depósito das patentes contribuiu significativamente para a escalabilidade e desempenho da estratégia, tornando o tempo de resposta na fase de busca bastante eficiente, mesmo diante de conjuntos de dados com número expressivo de registros, como adotado neste estudo com mais de 800 mil documentos.

Ademais, os resultados obtidos evidenciaram a relevância semântica dos resultados da busca, mesmo diante da ausência de uso de palavras-chave. A capacidade do modelo de considerar sinônimos, variações morfológicas e relações semânticas contextuais se demonstrou relevante para a recuperação de patentes pertinentes, enriquecendo o escopo de documentos encontrados sem comprometer a precisão. Enfatizando que a adoção de um limite mínimo de similaridade (30%) foi uma medida eficaz para remover as patentes irrelevantes, contribuindo para a pertinência dos resultados apresentados.

Dessa forma, a estratégia proposta neste estudo representa uma contribuição relevante para o campo da organização e recuperação da informação, propriedade industrial e ciência da informação, permitindo que pesquisadores e analistas realizem buscas mais intuitivas através do uso da linguagem natural.

BIBLIOGRAFIA

Amadei, J. R. P., & Torkomian, A. L. V. (2009). As patentes nas universidades: análise dos depósitos das universidades públicas paulistas (1995-2006). Ciência da Informação, 38(2), 9–18. https://repositorio.usp.br/item/001772714

Barros, T. S. (2022). Um Modelo BERT para Sumarização Extrativa de Textos em Documentos da Polícia Federal [Dissertação de mestrado]. Biblioteca Digital de Teses e Dissertações da UFCG. https://dspace.sti.ufcg.edu.br/handle/riufcg/27174

Brandão, F. G. (2016). Democratização da informação a partir do uso de repositórios digitais institucionais: da comunicação científica às informações tecnológicas de patentes [Dissertação de Mestrado]. UFRGS LUME. https://lume.ufrgs.br/handle/10183/179853

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [v.2]. ArXiv. 10.48550/arXiv.1810.04805

Instituto Nacional da Propriedade Industrial. (2024). Boletim mensal de propriedade industrial: Estatísticas preliminares. Instituto Nacional da Propriedade Industrial. https://www.gov.br/inpi/pt-br/central-de-conteudo/estatisticas/arquivos/publicacoes/boletim-mensal-de-pi_resultados-de-setembro-2024_v-2.pdf

Lei n. 9.279, de 14 de maio de 1996. (1996). Presidência da República.http://www.planalto.gov.br/ccivil_03/leis/l9279.htm

Liu, Y. (2019). Fine-tune BERT for Extractive Summarization [v2]. ArXiv. 10.48550/arXiv.1903.10318

Lopes, I.L. (2002). Estratégia de busca na recuperação da informação: revisão da literatura. Ciência da Informação, 31(2), 60-71. https://doi.org/10.18225/ci.inf.v31i2.961

Manglano, B. G., & Zulueta, M. A. (2007). Estudio comparativo de bases de datos de patentes en Internet. Anales de Documentación, 10, 145-162. https://revistas.um.es/analesdoc/article/view/1121

Nascimento, M. G., & Speziali, R. S. (2020). Patentometria: a utilização de dados contidos em patentes como mecanismo de análise da predominância tecnológica dos nits [Comunicação oral]. IV Encontro Internacional de Gestão, Desenvolvimento e Inovação.

Pires, E. A., Ribeiro, N. M., & Quintella, C. M. (2020). Sistemas de Busca de Patentes: análise comparativa entre Espacenet, Patentscope, Google Patents, Lens, Derwent Innovation Index e Orbit Intelligence. Cadernos De Prospecção, 13(1), 13. https://doi.org/10.9771/cp.v13i1.35147

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. ArXiv. https://doi.org/10.48550/arXiv.1908.10084

Reymond, D., & Quoniam, L. (2018). Patent documents in stem and phd education: Open-source tools and some examples to open discussion. In EDUCON 2018 IEEE Global Engineering Education Conference, 2018, Spain (pp. 4–9). 10.1109/EDUCON.2018.8363100

Rezende, N., Dalip, D., Brandão, M., & Vasconcelos, M. (2023). Elaboração de um Conjunto de Dados sobre o Registro de Patentes no Brasil. In C. F. Dornelles, E. Araújo, & M. M. Moro (orgs.), Anais do V Dataset Showcase Workshop, 2023, Brasil (pp. 99-108). Sociedade Brasileira de Computação. https://doi.org/10.5753/dsw.2023

Sanz-Casado, E. (2006). Los estudios métricos de la información y la evaluación del a actividad científica: conceptos básicos [Material didático do curso “Os estudos métricos da informação”, ministrado no Programa de Pós-graduação em Ciência da Informação da ECA/USP].

Escola de Comunicação e Artes da Universidade de São Paulo.

Silva, R. R., Dias, T. M. R., & Segundo, W. L. R. C. (2024). Protocolo de coleta de dados bibliográficos de patentes. Encontro Brasileiro de Bibliometria e Cientometria, 9(1), 1-8. https://doi.org/10.22477/ix.ebbc.361

World Intellectual Property Organization. (2022). Com crescimento impulsionado pela Ásia, depósitos de PI em todo o mundo registram novo recorde histórico em 2021. WIPO. https://www.wipo.int/pressroom/pt/articles/2022/article_0013.html

ANEXO 1

RESUMO BIOGRÁFICO DOS AUTORES

Raulivan Rodrigo da Silva

Doutorando em Modelagem Matemática e Computacional pelo CEFET-MG (2022). Mestre em Modelagem Matemática e Computacional pelo CEFET-MG (2022). Atualmente sou professor efetivo do CEFET MG no campus Divinópolis, lotado no Departamento de Informática, Gestão e Design (2019). ORCID: 0000-0002-2740-1045

Thiago Magela Rodrigues Dias

Doutor em Modelagem Matemática e Computacional pelo CEFET-MG (2016) tendo trabalhado com Bibliometria, Extração de Dados Científicos e Análise de Redes de Colaboração Científica. Mestre em Modelagem Matemática e Computacional pelo CEFET-MG (2008). Possui graduação em Ciência da Computação pelo Centro Universitário de Formiga - UNIFOR (2004), além de Especialização em Produção de Software - com Ênfase em Software Livre pela UFLA (2007) e Especialização em Melhoria do Processo de Software, UFLA (2007). Atua como Professor no Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG). ORCID: 0000-0002-3090-9413

Washington Luís Ribeiro de Carvalho Segundo

É Doutor e Mestre em Informática pela Universidade de Brasília, com Estágio de Doutorado Sanduíche no King's College London. Possui graduação em Matemática (Bacharelado e Licenciatura) também pela Universidade de Brasília. É Coordenador Técnico da Área de Tratamento, Análise e Disseminação da Informação Científica no Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict / MCTI). ORCID: 0000-0003-3635-9384