10° EBBC
Encontro Brasileiro de Bibliometria e Cientometria

UFPR
UNIVERSIDADE FEDERAL DO PARANÁ

OTIMIZAÇÃO DA RECUPERAÇÃO DE INFORMAÇÃO TECNOLÓGICA:
Uma Abordagem de Busca Semântica em Patentes com a Ferramenta Pytente

Raulivan Rodrigo da Silva
https://orcid.org/0000-0002-2740-1045
raulivan@cefetmg.br
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)

Thiago Magela Rodrigues Dias
https://orcid.org/0000-0001-5057-9936
thiagomagela@cefetmg.br
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)

Washington Luís Ribeiro de Carvalho Segundo
https://orcid.org/0000-0003-3635-9384
washingtonsegundo@ibict.br
Instituto Brasileiro de Informação em Ciência e Tecnologia (Ibict)

Fábio Pereira
https://orcid.org/0009-0006-4733-9484
fabiopereiragt@gmail.com
Universidade Federal de Lavras (UFLA)

Resumo:
Este trabalho apresenta um método para a recuperação de informações tecnológicas em documentos de patentes por meio de busca semântica. A pesquisa utiliza uma ferramenta computacional para coleta de dados e modelos de aprendizado profundo para identificar conceitos em vez de palavras isoladas. O experimento processou uma base de dados brasileira com milhares de registros, demonstrando agilidade e precisão na identificação de patentes. Os resultados indicam que a similaridade conceitual permite filtrar ruídos de forma eficiente. Conclui-se que a estratégia facilita a prospecção tecnológica ao superar as complexidades da linguagem técnica, tornando a análise de patentes mais acessível.

Palavras-Chave: Patente. Pytente. Busca semântica.

EIXO TEMÁTICO:
Ciência, Tecnologia e Inovação

MODALIDADE:
Resumo expandido

DOI: 10.22477/x.ebbc.1128

COMO CITAR ESTE ARTIGO:

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; SEGUNDO, Washington Luís Ribeiro de Carvalho; PEREIRA, Fábio. Otimização da recuperação de informação tecnológica: uma abordagem de busca semântica em patentes com a ferramenta Pytente. *In*: EBBC - Encontro Brasileiro de Bibliometria e Cientometria. 10. **Anais [...]**. Curitiba, 2026. DOI: 10.22477/x.ebbc.1128. Disponível em: https://doi.org/10.22477/x.ebbc.1128.

1 INTRODUÇÃO

O campo da Propriedade Intelectual exerce um papel fundamental no fomento à inovação e no desenvolvimento tecnológico. Nesse cenário, as patentes consolidam-se como fontes primárias de informação técnica, revelando tendências de mercado e estratégias competitivas de organizações globais. Diferentemente de outros registros, o documento de patente exige descrições exaustivas das invenções, abrangendo diagramas técnicos, esquemas ilustrativos e reivindicações que delimitam o escopo de proteção legal da tecnologia (Cativelli; Pinto; Sanchez, 2023). Tal nível de detalhamento torna esses documentos recursos essenciais para pesquisadores. Conforme apontam Reymond e Quoniam (2018), a densidade de informações contida em bases de patentes é substancialmente superior àquela encontrada em artigos científicos tradicionais, o que amplia sua relevância como instrumento de prospecção e análise tecnológica.

Dados da Organização Mundial da Propriedade Intelectual (OMPI) corroboram essa importância ao estimar que aproximadamente 70% das informações divulgadas em patentes não são publicadas em qualquer outro tipo de literatura (INPI, 2020). Instituições como a Organização para a Cooperação e Desenvolvimento Econômico (OCDE) e fóruns especializados, como o Fórum Nacional de Gestores de Inovação e Transferência de Tecnologia (FORTEC), reforçam a necessidade de explorar esses ativos (Silva; Nascimento, 2020).

Entretanto, a problemática central que motiva esta investigação reside no fato de que a extração de conhecimento desse tipo de documento enfrenta desafios significativos decorrentes da natureza intrínseca dos registros. A redação de patentes é caracterizada por uma linguagem híbrida, técnico-jurídica, muitas vezes obscura para pesquisadores que não detêm expertise em Propriedade Intelectual.

Nesse cenário, este estudo apresenta o emprego de Processamento de Linguagem Natural (PLN) como uma alternativa promissora para mitigar as barreiras impostas pela complexidade terminológica. A aplicação de técnicas de inteligência artificial permite a simplificação de termos e a extração de semântica encoberta nos textos, facilitando a recuperação de informações por similaridade de conceitos em vez de apenas palavras-chave literais.

Diferente da recuperação de informação baseada estritamente na correspondência exata de termos, a busca semântica propõe uma mudança de paradigma ao focar na intenção do usuário e no significado contextual das palavras. Essa abordagem utiliza estruturas de conhecimento e relações ontológicas para superar a superfície do texto, permitindo que o sistema identifique conceitos equivalentes mesmo quando expressos por vocabulários distintos.

Diante do exposto, o presente estudo tem como objetivo principal propor um método para a recuperação de dados bibliográficos de patentes utilizando técnicas de PLN. Para tanto, torna-se necessário o alcance do seguinte objetivo específico: a constituição de um repositório local, composto pelos metadados bibliográficos de patentes depositadas no Brasil. A estruturação deste corpus próprio viabiliza a aplicação de algoritmos de PLN.

2 METODOLOGIA

Este estudo caracteriza-se como uma pesquisa de natureza descritiva, voltada ao desenvolvimento e aplicação de novas estratégias para a recuperação de informações tecnológicas em conjuntos de patentes. A fundamentação do trabalho sustenta-se em uma abordagem documental e bibliográfica, abrangendo a literatura de Propriedade Industrial, Patentometria e Ciência da Informação.

O processo de aquisição de dados foi viabilizado pela ferramenta Pytente, uma solução computacional de acesso aberto desenvolvida para automatizar a coleta e armazenamento de informações bibliográficas de patentes. A ferramenta fundamenta-se no Protocolo CTDP (Protocolo de Coleta de Dados Bibliográficos Patentários), proposto por Silva, Dias e Carvalho Segundo (2024).

Conforme mencionado, para a constituição do corpus, utilizou-se a coleta sistemática de dados bibliográficos por meio da ferramenta Pytente. O escopo da investigação delimitou-se à recuperação de patentes depositadas no Brasil no intervalo compreendido entre 1900 e 2024, por meio do repositório internacional Espacenet (pd="19000101 20241231" AND pn="BR"). Os dados coletados foram armazenados em arquivos no formato ".json", adotando-se a estratégia de um arquivo por patente, com o objetivo de otimizar a organização e a manipulação do corpus. Adicionalmente, foi estabelecida uma convenção de nomenclatura padronizada, na qual o nome de cada arquivo corresponde ao número de depósito da respectiva patente.

Após a definição do corpus , o processamento semântico tem início com a (1) entrada de dados, com a extração do número de depósito, título e resumo das patentes coletadas. Em seguida, realiza-se a limpeza de dados, na qual são excluídos todos os registros que não contenham informações nos campos selecionados (número, título ou resumo). Por fim, é feita a 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 de depósito e a descrição.

Em seguida, é realizado o (2) pré-processamento, onde são realizadas etapas fundamentais para a preparação dos dados, como a Tokenização, que consiste na

divisão do texto em unidades menores chamadas tokens. Esses tokens incluem todas as palavras, pontuações e símbolos presentes no título ou resumo das patentes. Após a tokenização, ocorre a etapa de Remoção de Pontuação, na qual todos os tokens compostos exclusivamente por caracteres de pontuação são removidos, uma vez que não contribuem para a interpretação semântica do texto.

Por fim, ocorre a remoção de Stop Words, processo no qual são excluídas palavras de alta frequência e baixo valor informativo, como [a], [o] e [de]. A eliminação dessas palavras é importante para evitar impactos negativos no desempenho do modelo, visto que, assim como a pontuação, não agregam significado relevante ao contexto analisado.

Na sequência, é realizada a (3) representação textual, em que são aplicadas técnicas que permitem a interpretação dos tokens pelo modelo. Uma das abordagens adotadas é a métrica TF-IDF (*Term Frequency-Inverse Document Frequency*), que atribui pesos aos termos de um documento com base em sua frequência local e global. Dessa forma, termos recorrentes em um documento específico, mas raros no conjunto total, recebem pesos mais elevados, contribuindo para uma representação mais eficiente. Outra técnica utilizada é o Word2Vec, um algoritmo que converte palavras em vetores numéricos em um espaço de baixa dimensão, estabelecendo relações semânticas entre os termos.

A base computacional 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. Além disso, 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.

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. 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 coletada.

Para a implementação de busca semântica no conjunto local de patentes, foi adotado o modelo pré-treinado “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.

Por fim, para a execução da (4) busca semântica, o texto usado para a busca, por exemplo, “sistema de descongelamento para aparelho de refrigeração”, também passa por um pipeline de 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 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).

3 RESULTADOS

A aplicação da estratégia de busca semântica demonstrou eficácia na recuperação de documentos relevantes a partir de descrições textuais formuladas em lingua-

gem natural. O experimento foi conduzido sobre um corpus robusto, composto por 831.706 registros únicos de patentes, o que totaliza aproximadamente 1 gigabyte de dados textuais processados. O esforço computacional despendido nas fases de pré-processamento e geração de vetores de embeddings totalizou 204 horas de processamento em um ambiente dotado de 16 GB de memória RAM e processador Intel Core i5-6300U (2.40GHz). Esse procedimento resultou na estruturação de 51 índices distintos, segmentados por ano de depósito, garantindo que o custo computacional de indexação seja um investimento único, passível de reutilização ou atualizações pontuais.

No que tange ao desempenho em tempo de execução, a solução proposta demonstrou viabilidade para ambientes de pesquisa que exigem agilidade. Embora a recuperação envolva a varredura sequencial dos índices de embeddings, o algoritmo foi implementado para retornar resultados parciais em tempo real, mitigando a percepção de latência pelo usuário. Em média, a busca completa em toda a base histórica é concluída em menos de dois minutos, um tempo de resposta satisfatório considerando a densidade vetorial do corpus.

A análise qualitativa revelou que a relevância dos resultados é substancial, com as patentes recuperadas apresentando correspondência semântica direta com as consultas em linguagem natural. Essa coerência é evidenciada na **Tabela 1**, que apresenta uma amostragem dos registros com maiores índices de similaridade para o enunciado “Dispositivo ou sistema de refrigeração ou resfriamento”.

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

Fonte: Elaboração do autor

Os dados apresentados na tabela permitem observar que o modelo identifica com precisão patentes que, embora utilizem variações lexicais como “refrigerado” ou “refrigerante”, mantêm a identidade conceitual com o termo de busca. A estrutura da tabela contempla o percentual de similaridade, o número de depósito e os 200 caracteres iniciais da descrição, evidenciando a relação direta entre o escore calculado e a pertinência do documento.

A análise empírica dos dados recuperados indicou que registros com índice de similaridade inferior a 30% apresentam um baixo grau de relevância, sendo frequentemente classificados como ruídos informacionais que não abordam diretamente a tecnologia consultada. Portanto, a definição de um limite mínimo de 30% de similaridade mostra-se essencial para garantir a precisão do modelo em aplicações práticas de prospecção tecnológica.

4 CONSIDERAÇÕES FINAIS

As evidências apresentadas neste estudo demonstram que a integração de técnicas avançadas de Processamento de Linguagem Natural, especificamente por meio de modelos de Sentence-BERT, representa um avanço significativo para a recuperação de informações em bases de patentes no idioma português.

Ao converter descrições complexas em representações vetoriais, foi possível estabelecer uma ponte entre a intenção de busca do usuário em linguagem natural e o conteúdo técnico das patentes, eliminando a dependência exclusiva de palavras-chave literais ou uso de operadores booleanos.

Ademais, a aplicação da ferramenta Pytente comprovaram a eficiência ao automatizar a coleta e o processamento de grandes volumes de dados bibliográficos, permitindo que a busca semântica supere as barreiras impostas pela linguagem técnico-jurídica intrínseca aos registros de Propriedade Industrial.

REFERÊNCIAS

CATIVELLI, A. S.; PINTO, A. L.; SANCHEZ, M. L. L. Construction of a patent value index by measuring of indirect economic and technological value. **Transinformação**, v. 35, 2023. DOI: https://doi.org/10.1590/2318-0889202335e220019. Acesso em: 13 mar. 2026.

INPI . **Busca de patentes**. Brasil, 2020. Disponível em: https://www.gov.br/inpi/pt-br/assuntos/informacao/busca-de-patentes. Acesso em: 05 de mar. de 2026.

REYMOND, D.; QUONIAM, L. Patent documents in stem and phd education: Open-source tools and some examples to open discussion. **EDUCON 2018 IEEE Global Engineering Education Conference**, p. 4–9, 2018. DOI: https://doi.org/10.1109/EDUCON.2018.8363100. Acesso em: 08 mar. de 2026.

SILVA, Raulivan Rodrigo da; DIAS, Thiago Magela Rodrigues; CARVALHO SEGUNDO, Washington Luís Ribeiro de. Protocolo de coleta de dados bibliográficos de patentes. **Encontro Brasileiro de Bibliometria e Cientometria,** [*S. l.*], v. 9, p. 1–8, 2024. DOI: 10.22477/ix.ebbc.361. Disponível em: https://ebbc.inf.br/ojs/index.php/ebbc/article/view/361. Acesso em: 13 mar. 2026.

SILVA, M. G. S. R. da; NASCIMENTO. Patentometria: uma ferramenta indispensável no estudo de desenvolvimento de tecnologias para a indústria química. **Quim. Nova**, v. 43, n. 10, p. 1538–1548, ago. 2020. DOI: http://dx.doi.org/10.21577/0100-4042.20170620. Acesso em: 08 mar. de 2026.