Um framework para extração automática de informações em patentes farmacêuticas

Pablo Cecilio[1], Antônio Pereira[1], Felipe Viegas[2], Juliana Rosa[3], Washington Cunha[2], Fabiana Testa[4], Elisa Tuler[1], Leonardo Rocha[1]
[1]Universidade Federal de São João del-Rei, [2]Universidade Federal de Minas Gerais, [3]Universidade do Porto, [4]Universidade de São Paulo
(cecilio,antoniopereira)@aluno.ufsj.edu.br,(etuler,lcrocha)@ufsj.edu.br
(frviegas,washingtoncunha)@dcc.ufmg.br,jviegas@i3s.up.pt,fabtesta@fcfrp.usp.br

Abstract

The management of pharmaceutical patents often involves laborious manual searches due to the extensive details in documents on the invention's claims and methodology/results explanation. In order to address this challenge, we have proposed **PATopics**, a comprehensive framework designed to extract pertinent information from textual data within patents. **PATopics** utilizes this information to construct relevant topics, establish correlations with useful patent characteristics, and present the gathered insights through a user-friendly web interface. To evaluate the effectiveness of our framework, we conducted a study involving 4,832 pharmaceutical patents associated with 809 molecules patented by 478 companies. We analyzed the framework's performance based on the requirements of three user profiles: researchers, chemists, and companies. The results highlighted the practicality and usefulness of **PATopics** in the pharmaceutical domain, showcasing its ability to assist users from different backgrounds in navigating and extracting valuable insights from patent information.

Keywords: Natural Language Processing, Topic Modeling

1 Introdução

As patentes farmacêuticas são fruto de projetos de pesquisa desenvolvidos na academia e em empresas, depositadas e registradas por meio de extensos documentos textuais, em grandes repositórios para se garantir suas propriedades intelectuais [1, 2]. Empresas e pesquisadores precisam continuamente realizar consultas a esses repositórios para obter informações no que diz respeito à gestão de patentes. Esses profissionais precisam ler documentos extensos para obter informações simples ou mais detalhadas. A criação de ferramentas que auxiliem esses profissionais a encontrarem informações de forma mais rápida e prática se torna essencial.

A Modelagem de Tópicos (MT) é a tarefa de aprendizado de máquina que **automaticamente** extrai tópicos "implícitos" de uma coleção de documentos e atribui os tópicos mais prováveis para cada documento [5]. Neste trabalho, propomos o **PATopics**, um *framework* especialmente projetado para buscar automaticamente patentes farmacêuticas da Web e criar tópicos semânticos. O **PATopics** é capaz de identificar os principais tópicos abordados por essas patentes, correlacionando-os aos inventores e suas instituições e/ou empresas. Ele é composto por quatro blocos de construção principais, a saber, (i) representação dos dados, (ii) modelagem de tópicos, (iii) correlação dos tópicos com inventores, instituições e empresas e (iv) interface de sumarização.

Instanciamos o **PATopics** e fornecemos uma extensa análise considerando um conjunto de dados composto por documentos referentes a 4.832 patentes farmacêuticas referentes a 809 moléculas patenteadas por 478 instituições/empresas. Instanciamos o primeiro bloco aplicando algumas estratégias de pré-processamento: conversão para letras minúsculas, remoção de pontuação, acento e *stopwords* e uma abordagem de reconhecimento de entidade. Além disso, utilizamos os conceitos de CluWords [5], consideradas o estado da arte, para representar semanticamente esses dados. No segundo bloco, adotamos a estratégia de modelagem de tópicos NMF para inferir os diferentes tópicos do nosso conjunto de patentes. Para o terceiro bloco, propomos uma estratégia que consiste na manipulação das matrizes fornecidas pelo NMF, que permite correlacionar os temas descobertos patentes, seus inventores e suas instituições/empresas. Por fim, apresentamos uma proposta de interface visual que resume todas as informações geradas, destacando os principais tópicos obtidos e suas correlações. Analisamos extensivamente o **PATopics** sob a perspectiva de três perfis de usuário (i.e. (i) os acadêmicos e empregadores que trabalham com busca de patentes, (ii) os químicos e desenvolvedores de patentes, e (iii) empresas e indústrias que usam, compram ou aplicam a tecnologia de transferência de patentes), respondendo positivamente à duas questões de pesquisa (QP) são: **QP1:** *O PATopics é capaz de resumir as patentes farmacêuticas em tópicos coerentes?* **QP2:** *Os tópicos farmacêuticos trazem informações relevantes para auxiliar os profissionais?*

In: Workshop de Ferramentas e Aplicações (WFA 2023), Ribeirão Preto, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2023.

ISSN 2596-1683

2 PATopics

O **PATopics** é dividido em quatro blocos: (i) Representação de dados, (ii) Decomposição na modelagem de tópicos, (iii) Correlação entre entidades e (iv) Interface sumarizada.

2.1 Representação de dados

Nesta etapa, diversas estratégias de representação de dados podem ser utilizadas para representar a descrição textual de patentes farmacêuticas. O *framework* possui implementados quatro tipos de representação de dados: TF-IDF [4], TF-IDF com bigramas, CluWords [5], CluWords com bigramas (representação de dados que combina a representação CluWords explorando bigramas). Resumidamente, a representação TF-IDF é uma das formas mais tradicionais de representação de dados textuais. É um vetor de comprimento fixo onde cada índice representa uma palavra na coleção do vocabulário. Os bigramas são geralmente usados para enriquecer a representação dos dados, onde palavras compostas são incluídas como um elemento único no vocabulário da coleção. Exploramos a função *Phrases gemsim* para construir os bigramas. Para reduzir o número de combinações, ignoramos todos os bigramas com score $(word_a, word_b) < 0,5$, onde a função *score* retorna a porcentagem de coocorrência em documentos da coleção. A representação CluWords é uma representação de dados que incorpora informação semântica para enriquecer a informação textual. O método possui três etapas principais: (a) *Clustering* – explora a abordagem dos vizinhos mais próximos para capturar o parentesco semântico; (b) Filtragem – filtra possíveis ruídos na vizinhança semântica; (c) Ponderação – combina a representação TF-IDF com a vizinhança semântica por meio de ponderação. Na instanciação apresentada neste trabalho consideramos as CluWords com bigramas.

2.2 Decomposição na modelagem de tópicos

Nesta etapa, o **PATopics** explora o método de modelagem de tópicos chamado *Non-negative Matrix Factorization* (NMF), que corresponde a uma fatoração de matriz onde uma matriz de entrada A é decomposta em duas matrizes $H \in \mathbb{R}^{n \times k}$ e $W \in \mathbb{R}^{k \times m}$ [3]. O objetivo é encontrar uma aproximação k que satisfaça $A \approx H \times W$. Cada k-dimensão é representada como um tópico no NMF. A matriz H codifica a relação entre os documentos e os tópicos (k-dimensionados), enquanto a matriz W codifica a relação entre as palavras e os tópicos.

2.3 Correlação entre entidades

Consideramos como entrada a coleta de dados com descrição textual de patentes farmacêuticas e as matrizes H e W decompostas pelo método NMF. Seguindo o exemplo, considere que a patente i^{th} da matriz H trata principalmente do tópico "Tratamento do câncer", enquanto a patente j^{th} trata de "Autoimune". Para este exemplo, cada patente possui um ou mais inventores, então é possível destacar quais tópicos são mais relacionados aos inventores por meio das relações entre patentes e tópicos encontrados. Da mesma forma, como os inventores de uma patente trabalham para empresas de pesquisa, também é possível destacar as empresas por tópicos, considerando a relação entre patentes e tópicos. A estratégia consiste em manipular as matrizes fornecidas pelo NMF que correlacionam tópicos e patentes, introduzindo informações dos inventores e suas instituições/empresas, conforme exemplo a seguir - considerando as matrizes H e W para três tópicos. Primeiro, cada tópico é identificado analisando a matriz H e descobrindo quais palavras são mais fortemente associadas a cada tópico. Assumindo o exemplo em que o primeiro tópico está associado principalmente a "Tratamento do câncer", o segundo a "Autoimune" e o terceiro a "Tratamento da dor". Analisando a matriz W que relaciona documentos e tópicos, tomando como exemplo a primeira matriz da Tabela 1, que contém três patentes, onde cada posição apresenta a "relevância" do tópico para o documento. Assim, agrupar e somar os valores dos tópicos obtidos para patentes pertencentes ao mesmo inventor nos leva à segunda matriz da Tabela 1. Assumindo que as três patentes da primeira matriz pertencem ao primeiro inventor da segunda matriz, inferindo a "relevância" de cada tópico para este inventor.

(a) NMF Resultante

	Cancer treat.	Autoimmune	Pain treat.
Patent 1	30	70	10
Patent 2	20	65	40
Patent 3	17	80	8

⇓

(b) Pertinência do inventor por tópico

	Cancer treat.	Autoimmune	Pain treat.
Inventor 1	67	215	58
Inventor 2	47	150	18
Inventor 3	20	65	40

⇓

(c) Pertinência normalizada do inventor por tópico

	Cancer treat.	Autoimmune	Pain treat.
Inventor 1	20%	63%	17%
Inventor 2	22%	70%	8%
Inventor 3	16%	52%	32%

Table 1. Cálculo de contribuições do Inventor para tópicos.

2.4 Interface sumarizada

O *framework* possui uma interface de visualização intuitiva que resume os tópicos, suas associações com inventores/empresas e as principais moléculas envolvidas[1]. Na página inicial, temos acesso ao número de patentes, empresas, moléculas relacionadas e o número de inventores que reivindicam uma invenção patenteável. A aba *Topics* detalha os tópicos identificados de acordo com as principais palavras relacionadas aos mesmos. O usuário pode definir quantos tópicos deseja visualizar, bem como quantas palavras por tópico. Cada tópico também pode ser intitulado pelo usuário, que nesse caso deve ser uma pessoa com conhecimento de assuntos farmacêuticos. É possível clicar em cada tópico, acessar as patentes abrangidas por ele e até mesmo acessar cada patente individualmente. A aba *Companies* é construída com base nos

[1]https://labpi.ufsj.edu.br/patopics/ (username: user-test - password: avaliacao)

tópicos obtidos e pode ser ajustada de acordo número de empresas expositoras por tópico para 5, 10, 15 ou 20. Clicando em uma determinada Empresa acessam-se seus os dados, que compreendem o número de patentes por tópico, o número e o título da cada patente. A aba *Molecules* é construída com base nas substâncias mais patenteáveis de cada tópico. É possível observar a porcentagem de cada molécula em cada uma e acessar os dados da molécula, que compreende as patentes relacionadas em cada tópico. Por uma questão de limitação de espaço, ilustramos na Figura 1 apenas a aba *Topics*.

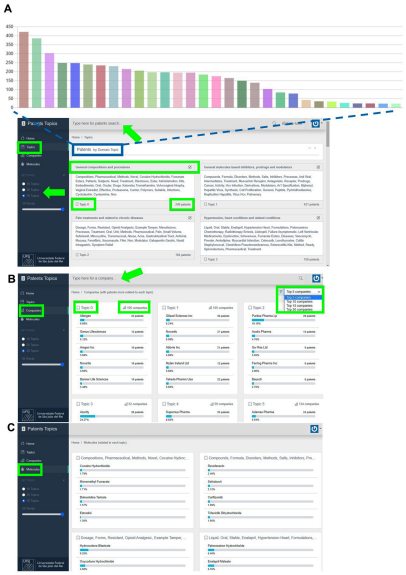

Figure 1. Ilustração do PATopics.

3 Análises e Discussões do PATopics

3.1 Coleta e limpeza de dados

Para avaliar o **PATopics** consideramos um conjunto de dados com 4.832 patentes farmacêuticas coletadas da plataforma WizMed[2]. Selecionamos as patentes escritas em inglês e publicadas entre 2003 a 2020 com as seguintes informações: **1**. Identificador de patente; **2**. Título da Patente; **3**. Descrição; **4**. Resumo; **5**. Molécula (substância); **6**. Empresa; **7**. URL; **8**. Dosagem; **9**. Nome comercial. Para construir a representação de dados descrita na Seção 2.1, exploramos os campos Título e descrição das patentes. Realizamos a remoção de *stopwords*, advérbios, verbos e intensificadores. Todos os resultados apresentados a seguir podem ser observados em https://labpi.ufsj.edu.br/patopics/(username:user-test-password:avaliacao)

3.2 Análise Geral

Primeiramente, analisando o total de patentes por ano, observamos que houve um aumento ao longo dos anos, atingindo um pico em 2014, porém com uma queda significativa no biênio 2020-2021, muito provavelmente pelo foco mundial na pandemia de Covid-19. O Tópico 1, relacionado à inibidores, pró-fármacos e moduladores baseados em moléculas gerais, é o tópico com maior número de patentes (421 patentes), seguido do Tópico 5, relacionado à métodos clínicos (385 patentes), e Tópico 6, relacionado à novos compostos e pró-fármacos (303 patentes). Alguns tópicos são altamente genéricos, como o Tópico 0 (249 patentes), Tópico 1 (421 patentes), Tópico 11 (248 patentes) e Tópico 21 (165 patentes), abrangendo uma gama de patentes que se correlacionam em algum ponto, mas pertencem a distintas áreas farmacêuticas. Alguns tópicos são bem mais específicos, como o Tópico 10, relacionado à condições e tratamentos dermatológicos, em que todas as 175 patentes são formulações farmacêuticas tópicas/transdérmicas para doenças de pele.

Avaliamos também a correlação entre as patentes e as instituições/empresas que patentearam, bem como as moléculas envolvidas nas patentes. *Allergan* e *Novartis* são as duas empresas que mais patenteiam (~130 patentes cada), seguidas pela *Takeda*. Essas três empresas detêm quase o dobro de patentes que as outras 10 maiores empresas. Cada um tem seus interesses específicos e o **PATopics** consegue identificar, por tópico, as empresas mais engajadas. Correlacionando os tópicos e as patentes, podemos observar, por exemplo, que os Tópicos 1 e 8 não apresentam o domínio de nenhuma empresa específica, tendo suas patentes distribuídas entre várias empresas. O tópico 10 mostra o domínio de Galderma, que detém 37 patentes neste tópico, seguido por 18 patentes da *Horizon*. O Tópico 29 ilustra um tópico no qual há uma parcela majoritária de patentes de propriedade de uma única empresa – 42 das 79 patentes pertencentes à *Amarin Pharma*. As patentes relacionadas à esses tópicos somam 410 patentes das quais o *Icosapent Ethyl* é a molécula mais patenteada.

Identificamos também os principais assuntos dessas patentes farmacêuticas. Os principais assuntos em ordem decrescente de representatividade: formulações e composições, novos compostos e pró-fármacos, condições crônicas, dor, métodos clínicos, dispositivos, vírus e câncer-relacionados, dermatológicos, gastrointestinais, terapia gênica, distúrbios cerebrais, oftálmicos e nasais. Como esperado, observamos muitas patentes relacionadas a novas formulações, composições, novos compostos e síntese de pró-fármacos, pois a maioria das patentes possui termos de formulação e composição em suas descrições. A estratégia de modelagem de tópicos consegue reunir as patentes de acordo com suas descrições. Outro tema mais patenteado cobre doenças crônicas, como hipertensão, doenças cardiovasculares e diabetes, onde estão os lucros substanciais das indústrias farmacêuticas [6].

3.3 Análise para Perfis Específicas

Do ponto de vista de um usuário pesquisador, o **PATopics** é capaz de fornecer acesso direto a patentes, podendo ser redirecionadas por meio de um link para o domínio em que foram depositadas. A ferramenta fornece funcionalidade para

[2](https://wizmed.com/drug-patent-database)

correlacionar patentes entre si. A possibilidade de trabalhar em um ambiente totalmente automático, onde as patentes já foram agrupadas em tópicos que as correlacionam, facilita o trabalho do pesquisador, otimizando o tempo de trabalho.

Os interesses comuns entre o usuário pesquisador e o usuário empresa e indústria são a possibilidade de realizar buscas rápidas, em ambiente padronizado, acompanhando atualizações de patentes existentes, acessando dados completos de patentes que vão desde seu texto original até dados sobre inventores, empresa responsável por patentes, moléculas ativas envolvidas, forma farmacêutica ou dispositivo, método de administração, dosagens e concentração das moléculas ativas.A ferramenta também apresenta análises que contribuem para a tomada de decisões sobre determinada pesquisa, uma possível compra ou o uso de uma determinada patente.

Uma empresa interessada em patentes, na utilização ou compra dos seus domínios, pode se beneficiar do **PATopics**, uma vez que o acesso a produtos relacionados permite uma comparação direta entre produtos, bem como obter informações das empresas envolvidas na patenteabilidade. O acesso a informações de patentes de empresas do mesmo nicho ajuda a desenvolver portfólios de produtos concorrentes no mercado. O acesso aos dados dos inventores pode contribuir diretamente para a contratação de funcionários focados em um determinado interesse de desenvolvimento. A ferramenta também pode auxiliar em estudos de linha do tempo, onde pode acompanhar tendências de patentes, auxiliando na implementação de processos de compra de domínios de produtos inovadores no mercado. É importante mencionar que entre as empresas e os desenvolvedores de patentes existe uma relação direta de ganho mútuo baseada na construção de redes, que possibilita uma parceria. Uma empresa nem sempre estará interessada em adquirir patentes. Em muitos casos, há uma dinâmica de negócios entre os dois lados, que os produz e utiliza para gerar lucros compartilhados.

Para esse terceiro usuário, desenvolvedores de patentes, geralmente químicos, o acesso a dados detalhados de patentes de forma simples auxilia no desenvolvimento de produtos no mesmo nicho e metodologias semelhantes. Atualmente, muitas patentes são dedicadas à síntese de pró-fármacos, e o acesso a essas patentes é de extrema importância para o químico responsável pelas sínteses. O conhecimento dos intermediários químicos utilizados, bem como das condições de síntese e do uso de catalisadores, ajudam a desenvolver metodologias cada vez mais otimizadas com altos rendimentos e pureza. É importante mencionar que, para esses desenvolvedores, o acesso às empresas envolvidas na patenteabilidade auxilia no processo de empregabilidade, onde a empresa é identificada por área de interesse e expertise. O ponto comum entre pesquisadores e desenvolvedores é o acesso à inovação, além da fácil identificação de moléculas e métodos altamente patenteáveis. As patentes farmacêuticas evoluem ao longo dos anos e, desta forma, uma formulação contendo uma determinada molécula, dispositivo de aplicação ou forma farmacêutica não continua a ser patenteada se não for altamente lucrativa. Inovações disruptivas estão cada vez mais presentes no nicho farmacêutico, e o acesso a esses dados em uma ferramenta, até onde sabemos, foi feito pela primeira vez no **PATopics**.

4 Conclusão e Trabalhos Futuros

As patentes farmacêuticas são compostas por documentos com muitos detalhes sobre a invenção e explicação da metodologia/resultados. Gerenciá-los corresponde a pesquisas manuais exaustivas. Para mitigar esse problema, propusemos o **PATopics**, um framework capaz de extrair informações relevantes de textos de patentes, construir tópicos relevantes, correlacioná-los com características úteis de patentes e apresentar as informações em uma interface web amigável. Avaliamos o framework usando 4.832 patentes farmacêuticas referentes a 809 moléculas patenteadas por 478 empresas. Nossas análises consideraram as demandas de três perfis de usuários – pesquisadores, químicos e empresas – mostrando a praticidade e utilidade do **PATopics** nesse cenário.

Para trabalhos futuros, há várias direções promissoras a serem exploradas. Primeiramente, iremos aprimorar a validação dos resultados obtidos pela ferramenta. Uma abordagem interessante seria calcular o percentual de trabalhos relevantes corretamente identificados em conjuntos específicos de testes, utilizando resultados previamente conhecidos para validação. Além disso, a avaliação manual de pesquisadores especializados na base de patentes poderia enriquecer essa validação. Paralelamente, permitir que os usuários forneçam *feedback* sobre os resultados e itens visualizados poderia ser uma forma eficaz de identificar falhas no algoritmo, permitindo que a equipe de moderação aprimore continuamente a ferramenta.

References

[1] Livio Garattini, Marco Badinella Martini, and Pier Mannuccio Mannucci. 2022. Pharmaceutical patenting in the European Union: reform or riddance. *Internal and Emergency Medicine* 17, 3 (2022), 937–939. https://doi.org/10.1007/s11739-021-02887-6

[2] B. L. Genin and D. S. Zolkin. 2021. Similarity search in patents databases. The evaluations of the search quality. *World Patent Information* 64, February (2021), 102022. https://doi.org/10.1016/j.wpi.2021.102022

[3] Zaiqiao Meng, Hong Shen, Huimin Huang, Wei Liu, Jing Wang, and Arun Kumar Sangaiah. 2018. Search result diversification on attributed networks via nonnegative matrix factorization. *Information Processing & Management* 54, 6 (2018), 1277–1291.

[4] Claude Sammut and Geoffrey I. Webb (Eds.). 2010. *TF–IDF*. Springer US, Boston, MA, 986–987. https://doi.org/10.1007/978-0-387-30164-8_832

[5] Felipe Viegas, Sérgio Canuto, Christian Gomes, Washington Luiz, Thierson Rosa, Sabir Ribas, Leonardo Rocha, and Marcos Gonçalves. 2019. CluWords: Exploiting Semantic Word Clustering Representation for Enhanced Topic Modeling. (2019). https://doi.org/10.1145/3289600.3291032

[6] Hugh Waters and Marlon Graf. 2018. The Costs of Chronic Disease in the U.S. *Milken Institute* August (2018), 24. https://milkeninstitute.org/sites/default/files/reports-pdf/ChronicDiseases-HighRes-FINAL.pdf