PATopics: Um framework para automatizar a extração de informações em documentos de patentes farmacêuticas*

Pablo Cecilio[1], Felipe Viegas[2], Juliana Rosa[3], Leonardo Rocha[1]

[1] Universidade Federal de São João del Rei, Brasil
[2] Universidade Federal de Minas Gerais, Brasil
[3] Universidade do Porto, Portugal
cecilio@aluno.ufsj.edu.br, jviegas@i3s.up.pt,
frviegas@dcc.ufmg.br, lcrocha@ufsj.edu.br

Abstract. *Pharmaceutical patents are composed of documents with many details regarding the invention's claims and methodology/results explanation. Management them refers to an exhaustive manual search. To mitigate this problem, we proposed* **PATopics***, a framework able to extract relevant information from patents' textual information, build relevant topics, correlate them with useful patent characteristics and present the information in a friendly web interface. We evaluated the framework using 4,832 pharmaceutical patents concerning 809 molecules patented by 478 companies. We analyze considering the demands of three user profiles – researchers, chemists, and companies – showing how practical and helpful* **PATopics** *is in the pharmaceutical scenario.*

Resumo. *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.*

1. Introdução

As patentes farmacêuticas desempenham um papel importante ao proteger a inovação de cópias, mas também impulsionam os pesquisadores a inovar, criar novos produtos e promover inovações disruptivas com foco na saúde coletiva [Garattini et al. 2022, Khachigian 2020]. Essas patentes, normalmente fruto de trabalhos desenvolvidos na academia, como os resultados de projetos de pesquisa, e em empresas, são depositadas e registradas, em grande parte com extensos documentos textuais, em grandes repositórios para se garantir suas propriedades intelectuais [Genin and Zolkin 2021]. Empresas e pesquisadores da área farmacêutica precisam continuamente realizar consultas a esses repositórios, além de realizar um trabalho manual exaustivo para obter informações no que diz respeito à gestão de patentes. Isso acontece porque os documentos de patente são complexos, com muitas informações e detalhes sobre a teoria por trás da invenção, explicação da

*Esse trabalho foi parcialmente financiado por AWS, CNPq, CAPES, FINEP e Fapemig

metodologia e detalhamentos dos resultados alcançados. Assim, às vezes, os profissionais precisam ler um documento extenso para obter informações simples ou mais detalhadas.

A Modelagem de Tópicos é um campo de pesquisa extenso na Ciência da Computação que ganhou muita atenção da comunidade científica nos últimos anos. A principal razão é a sua aplicabilidade em diversos contextos sociais e práticos. A Modelagem de Tópicos 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 [Viegas et al. 2019]. Neste trabalho, propomos o **PATopics**, um *framework* especialmente projetado para buscar automaticamente patentes farmacêuticas da Web e criar tópicos semânticos usando abordagens de Modelagem de Tópicos. Mais especificamente, 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. No primeiro bloco, a partir do conteúdo textual extraído das patentes, é possível aplicar diferentes estratégias de pré-processamento e de representação semântica de dados. A modelagem de tópicos, por sua vez, visa encontrar tópicos semânticos, ou temáticas, de artigos – várias estratégias podem ser utilizadas, tais como *Non-Negative Matrix Fatorization* (NMF) e LDA. Uma vez identificadas as principais temáticas das patentes, o terceiro bloco consiste em associá-las aos inventores, instituições e/ou empresas. Finalmente, a interface de sumarização fornece uma visualização intuitiva dos tópicos identificados e as associações estabelecidas.

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 [Viegas et al. 2019], 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?*

Todas as implementações e execuções dos experimentos foram realizadas pelo aluno Pablo Cecílio, sob a orientação do professor Leonardo Rocha. A concepção do projeto e as análises de resultados foram feitas em colaboração com o doutorando Felipe Viegas e a pesquisadora da Universidade do Porto, Juliana Rosa.

2. PATopics

O objetivo do *framework* proposto é construir tópicos semânticos para patentes farmacêuticas e usar esses tópicos para correlacioná-los com inventores, compostos químicos e empresas farmacêuticas. Dividimos a instanciação do **PATopics** em quatro etapas principais: (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 [Sammut and Webb 2010], TF-IDF com bigramas, CluWords [Viegas et al. 2019], 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) [Meng et al. 2018]. O método NMF é uma fatoração de matriz onde uma matriz de entrada A é decomposta em duas matrizes $H \in \mathbb{R}^{n \times k}$ e $W \in \mathbb{R}^{k \times m}$. O objetivo é encontrar uma aproximação k que satisfaça $A \approx H \times W$. Cada k-dimensão é representada como um tópico no método NMF. A matriz H codifica a relação entre os documentos e os tópicos (k-dimensonados), enquanto a matriz W codifica a relação entre as palavras e os tópicos.

2.3. Correlação entre entidades

Neste caso, 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 2.3, 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 Tabela2.3. 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%

Tabela 1. Cálculo das contribuições do Inventor para os tópicos.

Considerando a matriz W do método NMF, o mesmo processo pode ser aplicado a todos os inventores. Podemos calcular a distribuição entre os tópicos a que se referem as patentes de cada inventor. Ao normalizar as linhas que representam os inventores na segunda matriz da Tabela 2.3, é possível medir o impacto da pesquisa de cada inventor em cada tópico (Terceira matriz da Tabela 2.3). Esse processo é repetido para extrair a relevância das patentes das empresas para cada tópico.

2.3.1. 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 tópicos obtidos e pode ser ajustada de acordo número de empresas expositoras por tópico para 5, 10, 15 ou 20.

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

É possível clicar em uma determinada Empresa e acessar os dados desta empresa, 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*

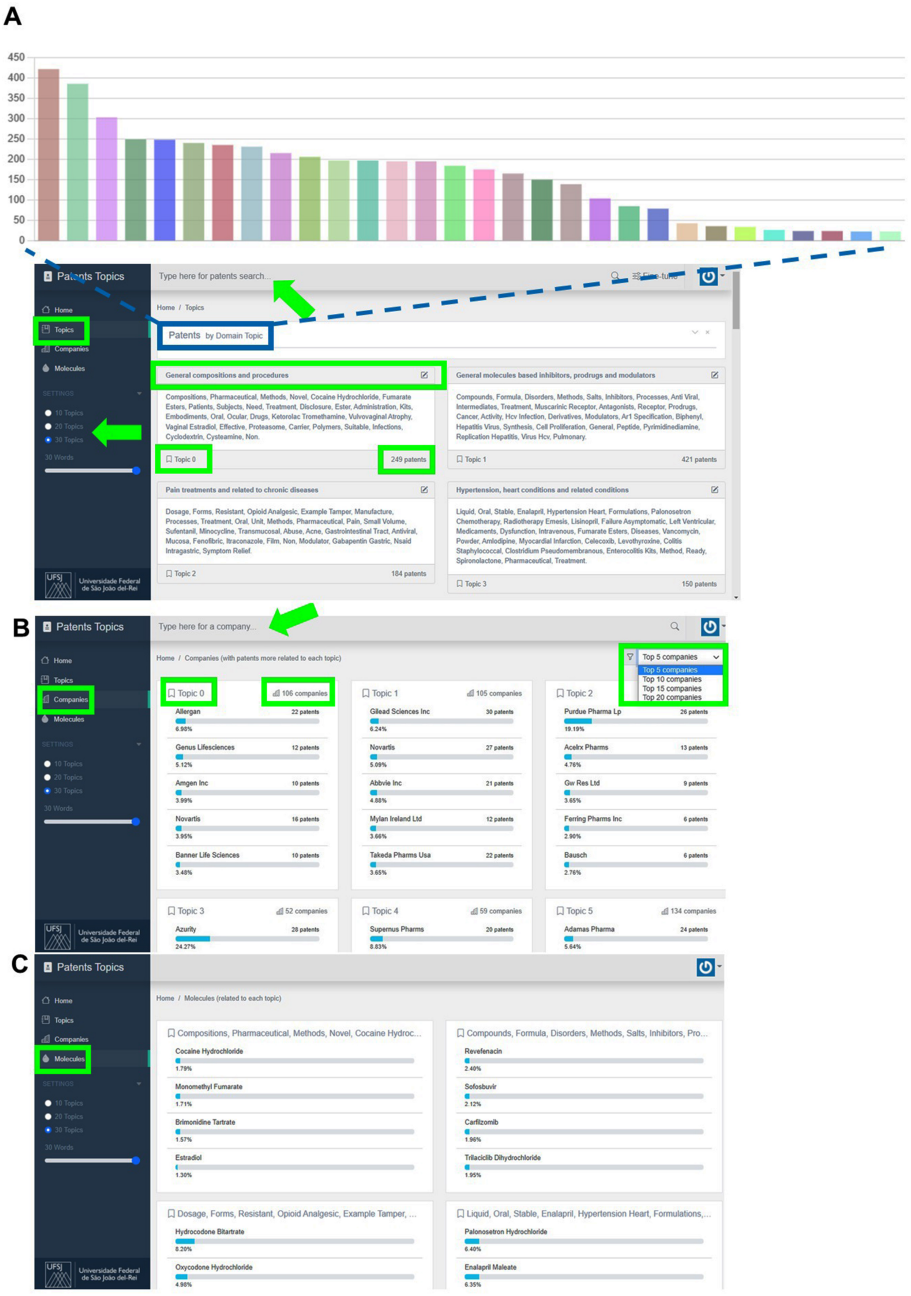

Figura 1. A seção Tópicos (A) contém uma barra de pesquisa e os tópicos gerados pelos grupos de palavras são descritos com possibilidade de título editável e o número de patentes por tópico é visível; A seção de empresas (B) possui uma barra de busca (por empresa) e elas são distribuídas nos tópicos gerados em 5, 10, 15 ou 20 empresas por tópico; A seção de moléculas (C) onde as moléculas mencionadas por tópico são destacadas.

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. Em termos de pré-processamento, realizamos a remoção de *stopwords* (utilizando o padrão SMART). Também removemos advérbios, verbos e intensificadores.

3.2. Análise Geral dos Tópicos, Moléculas e Instituições Correlacionadas

As análises apresentadas nessa seção foram realizadas e validadas em conjunto com a pesquisadora Juliana Rosa, pesquisadora no Instituto de Investigação e Inovação em Saúde na Universidade do Porto, e podem ser observadas na instanciação da ferramenta[3]. 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. Por exemplo, no Tópico 11, relacionado à formulações gerais, métodos e novas formas e dispositivos de administração, uma formulação em comprimido contendo um agente hipoglicemiante é agrupada com uma patente sobre uma cápsula de um analgésico. Isso acontece porque ambos estão relacionados a formulações (formas farmacêuticas), mas são de classes de medicamentos diferentes. De resto, 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 ($\sim$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 com base em patentes, 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

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

[3]`https://labpi.ufsj.edu.br/patopics/`. **username**: user-test - **password**: avaliacao

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, responsável por 67 patentes.

Identificamos também os principais assuntos dessas 4.832 patentes farmacêuticas e também as principais moléculas e empresas. 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. Não identificamos muitas patentes, como se esperava, em relação a vírus e câncer, mas talvez dados futuros, após a pandemia de Covid-19, esses dados mudem, dado o investimento em antivirais. No entanto, é importante mencionar os longos períodos de pesquisa necessários para a obtenção de novas moléculas e/ou formulações eficazes. O terceiro assunto mais patenteado cobre doenças crônicas, como hipertensão, doenças cardiovasculares e diabetes. De fato, este assunto concentra-se nos lucros substanciais das indústrias farmacêuticas [Reinhardt 2001, Waters and Graf 2018]. Cem milhões de americanos têm dislipidemia, um desequilíbrio do colesterol, indicando altos níveis de colesterol de lipoproteína de baixa densidade (LDL) ou baixos níveis de colesterol de lipoproteína de alta densidade (HDL). 80 milhões de pessoas vivem com hipertensão, que é um fator de risco significativo para doenças cardíacas e derrames. 10,9% da população adulta vivia com diabetes tipo 2. Hipertensão, dislipidemia e diabetes tipo 2, em soma, representaram cerca de 1,7 milhões de dólares de custos por ano nos EUA [Waters and Graf 2018]. Portanto, espera-se o aumento de patentes envolvendo produtos para doenças crônicas, bem como novas moléculas aplicadas a esse assunto.

3.3. Análise das Contribuições do PATopics para Diferentes Perfis

Nessa seção destacamos as principais contribuições do **PATopics**, tanto em um contexto geral, como também para diferentes perfis beneficiários da ferramenta: (a) Pesquisadores que trabalham com patentes, (b) Químicos que desenvolvem patentes e (c) Empresas ou indústrias que desejam usar ou comprar patentes.

3.3.1. Contribuições no Contexto Geral

Em relação às patentes farmacêuticas, uma das principais desvantagens é a descentralização dos dados de patentes ao redor do mundo. Cada país tem seu próprio repositório de patentes, ou mesmo as instituições dentro dos países depositam essas patentes por conta própria em seu próprio repositório. Dessa forma, quando há uma busca para consultar patentes, é desafiador encontrar patentes que representem fielmente a consulta. Não existem grandes repositórios de patentes estruturados para cobrir um grande número de patentes em diferentes tópicos. Por esse motivo, a principal contribuição geral do **PATopics** é sua capacidade de centralizar as patentes, a partir de uma consulta de busca utilizada para buscar as patentes, em um único ambiente de busca. Além disso, o **PATopics** não só reúne patentes, como também consegue agrupá-las em tópicos com base em suas semelhanças, o que facilita ainda mais a busca por tópicos. Este resumo

de tópicos de patentes na área farmacêutica consegue apontar os principais assuntos patenteados, destacar os tópicos mais relevantes, direcionar o interesse das empresas para o tópico científico, destacar as moléculas mais patenteadas bem como notificar a evolução das formas farmacêuticas, embalagens , dispositivos, métodos, entre outros. A ideia de reunir patentes de uma grande área, como a farmacêutica em um *framework*, facilita o acesso aos dados e possibilita a discussão e comparação de patentes desenvolvidas em diferentes lugares por diferentes empresas. A discussão e comparação, neste caso, permitem uma grande evolução do processo de criação e inovação, onde, através da ferramenta, é possível observar a evolução ao longo do tempo, expressa numericamente e facilmente interpretada. Desta forma, o **PATopics** representa não só uma ferramenta onde pode haver busca, mas também onde há análise de dados, similaridades, evolução na linha do tempo, crescimento do patenteamento por empresas e identificação de áreas e possibilidades.

3.3.2. Contribuições para Perfis Específicas

Nesta seção destaca as contribuições do **PATopics** para diferentes perfis de usuários com diferentes objetivos. A Figura 2 detalha os três perfis a serem analisados.

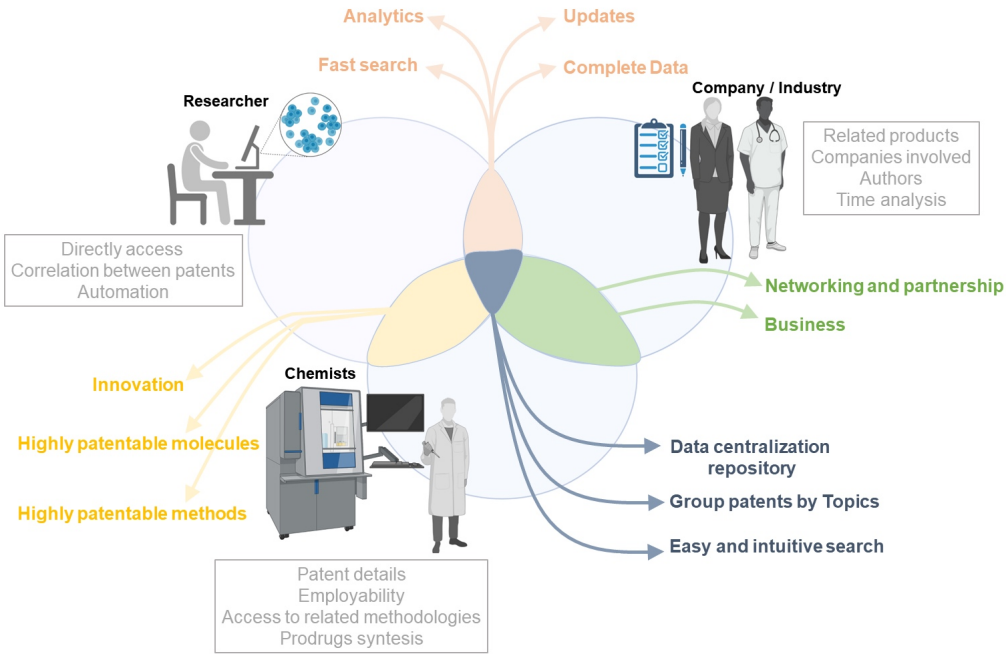

Figura 2. O perfil dos potenciais usuários do *framework* PATopics e seus principais interesses que o *framework* pode envolver. O primeiro perfil é de pesquisadores que trabalham com patentes e seus estudos. A segunda são os químicos, que desenvolvem as patentes e a terceira são as empresas e indústrias, que compram ou utilizam as patentes.

Do ponto de vista de um usuário pesquisador, o **PATopics** é um ambiente 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 correlacionar patentes entre si. Além disso, 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. Além disso, a ferramenta 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 beneficiar da utilização de uma plataforma como o **PATopics**, uma vez que o acesso a produtos relacionados permite uma comparação direta entre produtos, bem como obter informações das empresas envolvidas na patenteabilidade. Além disso, o acesso a informações de patentes de empresas do mesmo nicho ajuda a desenvolver portfólios de produtos concorrentes no mercado, bem como a inovação. 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

Neste trabalho, propusemos o **PATopics**, um *framework* especialmente desenvolvido para fornecer informações resumidas sobre patentes farmacêuticas. O **PATopics** é composto por quatro etapas de construção (i) Representação de dados, (ii) Decomposição de modelagem de tópicos, (iii) Correlação entre entidades e (iv) Interface sumarizada. Para representação de dados, o *framework* explora a representação CluWords combinada

com bigramas. O CluWords é a representação de dados de última geração que explora informações semânticas fornecidas por incorporações de palavras que enriquecem as informações textuais. Quanto à decomposição de modelagem de tópicos, o *framework* explora o NMF, uma das abordagens de modelagem de tópicos não probabilísticas mais consistentes na literatura. As etapas de correlação entre entidades é a etapa específica que associa os tópicos recuperados pelo NMF com características úteis sobre as patentes, como inventores, empresas e moléculas, enquanto a interface sumarizada é a interface web amigável que mostra todas as informações para o usuário final.

Em nossa avaliação experimental, instanciamos o **PATopics** usando uma amostra de 4.832 patentes farmacêuticas referentes a 809 moléculas patenteadas por 478 empresas extraídas da plataforma WizMed. Apresentamos como funciona o *framework*, mostrando todas as interfaces web do *framework* e quais informações podem ser extraídas delas. Além disso, apresentamos três perfis de usuários em potencial que podem tirar proveito do **PATopics**. O primeiro é o perfil do Pesquisador, que trabalha com patentes. O segundo perfil são os Químicos, que desenvolvem patentes, e por último, as Empresas que desejam utilizar ou comprar patentes. Nossa análise mostrou que o **PATopics** é útil para todos os três perfis, especialmente para pesquisadores e empresas, uma vez que esses dois perfis realizam mais pesquisas de patentes. Para trabalhos futuros, pretende-se melhorar a aplicação aprimorando a etapa de decomposição da modelagem de tópicos. Pretendemos construir um tópico hierárquico de patentes. O fator hierárquico poderia atenuar a geração de tópicos genéricos e trazer informações adicionais para entidades de correlação (terceira etapa), pois as informações hierárquicas poderiam ser utilizadas como um componente adicional para patentes relacionadas em camadas distintas da composição dos tópicos.

Referências

Garattini, L., Badinella Martini, M., and Mannucci, P. M. (2022). Pharmaceutical patenting in the European Union: reform or riddance. *Internal and Emergency Medicine*, 17(3):937–939.

Genin, B. L. and Zolkin, D. S. (2021). Similarity search in patents databases. The evaluations of the search quality. *World Patent Information*, 64(February):102022.

Khachigian, L. M. (2020). Pharmaceutical patents: reconciling the human right to health with the incentive to invent. *Drug Discovery Today*, 25(7):1135–1141.

Meng, Z., Shen, H., Huang, H., Liu, W., Wang, J., and Sangaiah, A. K. (2018). Search result diversification on attributed networks via nonnegative matrix factorization. *Information Processing & Management*, 54(6):1277–1291.

Reinhardt, U. E. (2001). Perspectives on the pharmaceutical industry. *Health Affairs*, 20(5):136–149.

Sammut, C. and Webb, G. I., editors (2010). *TF–IDF*, pages 986–987. Springer US, Boston, MA.

Viegas, F., Canuto, S., Gomes, C., Luiz, W., Rosa, T., Ribas, S., Rocha, L., and Gonçalves, M. A. (2019). Cluwords: Exploiting semantic word clustering representation for enhanced topic modeling. pages 753–761.

Waters, H. and Graf, M. (2018). The Costs of Chronic Disease in the U.S. *Milken Institute*, (August):24.