TÉCNICAS DE MINERAÇÃO DE TEXTO APLICADAS EM ANÁLISE DE PATENTES

Letícia Silveira Artese[1]

Luciano Zamperetti Wolski[2]

Rafael Sanceverino Mattos[3]

Alexandre Leopoldo Gonçalves[4]

José Leomar Todesco[5]

Abstract. *The high volume of patent data requires tools that can automate the endeavor in decision-making. This article apply text mining techniques and tools to extract and cluster terms related to nanotechnology from the US patent base USPTO, regarding 2019. The data collected were pre-processed, stored and indexed in order to use a search engine. Cluster analysis and visualization were also carried out. Thus contributing to elucidate the text mining process and exhibit the scope of the term nanotechnology on patent database.*

Keywords: Text Mining; Patent Analysis; Patent Mining; Nanotechnology.

Resumo. *O volume de dados gerado por patentes requer o uso de ferramentas que possam automatizar o esforço na tomada de decisão. Este artigo tem como objetivo demonstrar a aplicação de técnicas e ferramentas de mineração de texto, para extrair e gerar agrupamentos, a partir da base de patentes americanas USPTO. Os dados de patentes são referentes ao ano de 2019 e relativos à área de nanotecnologia. Os dados coletados foram pré-processados, armazenados e indexados para utilização em uma ferramenta de busca. Foi também realizada a analise e visualização dos agrupamentos. Contribuindo, assim, para a elucidação do processo de mineração de texto e exibindo a abrangência do termo nanotecnologia na base de patentes.*

Palavras-chave: Mineração de texto; Análise de patentes; Mineração de Patentes; Nanotecnologia.

[1] Doutoranda do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (PPGEGC) – Universidade Federal de Santa Catarina (UFSC), Florianópolis – Brasil. Correio electrônico: artese.leticia@gmail.com

[2] Doutorando do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (PPGEGC) – Universidade Federal de Santa Catarina (UFSC), Florianópolis – Brasil. Correio electrônico: lwolski@gmail.com

[3] Mestrando do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (PPGEGC) – Universidade Federal de Santa Catarina (UFSC), Florianópolis – Brasil. Correio electrônico: rsmattos@gmail.com

[4] Prof. Dr. do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (PPGEGC) – Universidade Federal de Santa Catarina (UFSC), Florianópolis – Brasil. Correio electrônico: a.l.goncalves@ufsc.br

[5] Prof. Dr. do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (PPGEGC) – Universidade Federal de Santa Catarina (UFSC), Florianópolis – Brasil. Correio electrônico: tite@egc.ufsc.br

1. INTRODUÇÃO

O conhecimento aplicado no desenvolvimento de soluções tecnológicas é consolidado com o depósito de uma patente. Processos de depósito de patentes tem como propósito conferir títulos de propriedade intelectual sobre determinada invenção concedendo o direito de exploração exclusiva por um determinado prazo (INPI, 2013). De fato, este é um importante papel da patente. Apesar disso, as informações contidas nos documentos de patentes conferem a possibilidade de explorá-la para além da sua vantagem econômica (Singh, Chakraborty, & Vincent, 2016).

A patente pode ser vista como fonte de informações contendo resultados tecnológicos raramente replicados em outras publicações, pois o pedido de patenteamento deve se tratar de algo novo, inédito, sem anterioridade (Singh et al., 2016). Portanto, um banco de patentes incluem informações valiosas que podem ser usadas, por exemplo, para identificação de tendências e padrões para auxiliar nas tomadas de decisões. Para Ernst (2003) as informações nos dados de patentes desempenham papel fundamental para fins de planejamento estratégico. Da mesma forma, para Madani e Weber (2016) os documentos de patentes contêm importantes resultados de pesquisa, pois representa uma invenção, um produto ou um processo, que fornece uma nova solução tecnológica para resolver um problema. No entanto, eles são longos e apresentam terminologia técnica, de modo que se faz necessário um grande esforço humano para analisá-los (Tseng et al., 2007). Dessa forma, a busca por ferramentas que possam automatizar essa etapa da análise de patentes, auxiliando nas tomadas de decisão, vem sendo cada vez mais requisitadas.

A análise de patentes é um campo de pesquisa composto por técnicas e ferramentas com a finalidade de obter informações contidas e relacionadas aos documentos de patentes. A análise envolve uma série de etapas, incluindo a extração dos documentos a partir dos bancos de dados de patentes, a extração das informações contidas nas patentes e a análise dessas informações resultando em inferências lógicas (Singh et al., 2016). Abbas, Zhang, e Khan (2014) classificam as técnicas de análise de patentes em técnicas de mineração de texto e técnicas de visualização. Essas técnicas e ferramentas são capazes de analisar e prever tendências tecnológicas, conduzir planejamento estratégico de tecnologia, detectar violação de patentes, determinar a qualidade

das patentes e as patentes mais promissoras, identificar *hotspots* tecnológicos e vacuidades patenteadas (Abbas, Zhang, & Khan, 2014). Portanto, a análise das patentes em um domínio específico confere apoio sob vários aspectos.

O domínio da nanotecnologia tem se mostrado promissor no século XXI dando indícios de muitos benefícios para a sociedade como um todo a partir de suas aplicações (Bayda, Adeel, Tuccinardi, Cordani, & Rizzolio, 2020). O campo da nanotecnologia diz respeito a manipulação e controle ao nível atômico e molecular, envolvendo conhecimentos provenientes de diferentes áreas como física, química, medicina, biotecnologia entre outros (Royal Society, 2004). Em 2000, a nanotecnologia foi proclamada uma prioridade para as pesquisas nos Estados Unidos, fundando a *National Nanotechnology Initiative* (NNI) (Porter et al., 2019), exemplificando, a importância, o interesse e o investimento na área. O campo apresenta uma relação próxima tanto com avanço na ciência quanto com o desenvolvimento tecnológico (Gao, Ding, Teng, & Pang, 2012), favorecendo o uso de base de patentes como fonte de investigação.

Verificada a extensão do tema nanotecnologia, o volume de dados que o campo abarca é considerável. O interesse em obter informações sobre o domínio está em sua relevância para a pesquisa, desenvolvimento e aplicação em projetos variados, garantindo a proficiência da área, minimizando retrabalho e propulsionando descobertas e inovações para o campo evidenciando as relações (NNI, 2012). A área de *Nanoinformatics* é responsável por desenvolver e implementar mecanismos eficazes para a comunidade de nanotecnologia coletar, validar, armazenar, compartilhar, minerar, analisar, modelar e aplicar informações acerca da nanotecnologia (Panneerselvam & Choi, 2014; Barnard et al., 2019; Afantitis, 2020). Afantitis, (2020) enfatiza a necessidade de análises que façam uso de técnicas e ferramentas computacionais, orientados à análise de dados, que auxiliem a fornecer uma infraestrutura de conhecimento para a área de nanotecnologia, baseada na comunidade e orientada a soluções auxiliando as pesquisas.

Estudos mais recentes, a fim de explicitar o domínio da nanotecnologia, se mostram segmentados em áreas específicas, como medicina (Ma, Abrams, Porter, Zhu, & Farrell, 2019), energia (Li et al., 2020), alimento e agricultura (Kah, Tufenkji, & White, 2019), nanocelulose (Charreau, Cavallo, & Foresti, 2020) entre outros, demonstrando uma escassez a respeito de uma visão geral da área. É fato que o domínio da nanotecnologia é amplo e multidisciplinar

(Stopar, Drobne, Eler, & Bartol, 2016), sendo desafiadora uma interpretação unificada desse amplo corpo científico de conhecimentos.

Como mencionado anteriormente, procedimentos para recuperar e representar informações presentes nas patentes são de grande valia pela ampla gama de aplicações desse conhecimento. A análise de patentes possibilita a extração de informações a partir de dados estruturados e não estruturados (Tseng et al., 2007). Com o propósito de explorar um domínio por meio da análise de patentes várias abordagens podem ser aplicadas (Shalaby & Zadrozny, 2019). Uma delas se dá pela investigação da similaridade entre patentes (Barnard et al., 2019), convertendo o texto da patente em um dado estruturado usando técnicas de mineração de texto (Antons, Grünwald, Cichy, & Salge, 2020).

A mineração de texto reúne técnicas que visam descobrir automaticamente, principalmente, padrões e tendências extraindo informações de grandes coleções de documentos não estruturados, textos. O campo da mineração de textos tem como fundamento abordagens da Recuperação da Informação (IR), Processamento de Linguagem Natural (NLP) e Descoberta de Conhecimento em Textos (KDT) viabilizando a descoberta de informações desconhecidas, ocultas (Fayyad & Uthurusamy, 1996). Pode ser compreendida como todo o processo desde o pré-processamento, passando pela mineração com a aplicação de algoritmos até o pós-processamento. Uma de suas principais aplicações ocorre pela identificação de padrões escondidos nos dados por meio da técnica de análise de agrupamentos ou c*lustering.* A análise dos agrupamentos permite identificar os documentos que sejam mais semelhantes entre si do que entre outros grupos (Allahyari et al., 2017). Os agrupamentos, desse modo, tornam possível a visualização, organização e classificação dos documentos.

Diante desse contexto, surgem as seguintes perguntas de pesquisa que norteiam este estudo: como as patentes associadas ao campo da nanotecnologia se organizam na base de patentes? Qual a abrangência do tema na base de patentes?

Isto posto, apresenta-se neste estudo um processo de mineração de texto na base de patentes a fim de realizar uma análise exploratória, gerando agrupamentos de patentes que tratem de temas correlatos, para investigar como o termo nanotecnologia se organiza na base de patentes. Neste sentido, objetiva-se apresentar um panorama dos tópicos relevantes para o

domínio da nanotecnologia referente ao ano de 2019 na base de patentes americanas, *United States Patent and Trademark Office* (USPTO).

2. MÉTODO

Com a finalidade de atender ao questionamento levantado por esta pesquisa, o seguinte método foi estabelecido, descrito na Figura 1.

Figura 1 - Etapas do método proposto.

Fonte: Os autores (2020)

A primeira etapa, se ocupa do referencial teórico que embasa a pesquisa. Mediante a consultas na literatura foram pesquisadas as definições e trabalhos na área de patentes, análise de patentes, nanotecnologia e mineração de texto, tópicos centrais a este estudo.

Na segunda etapa, foi selecionada a base de dados para a aplicação da pesquisa. Segundo (Singh et al., 2016) os principais repositórios para depósito de patente são: *United States Patent and Trademark Office* (USPTO), *European Patent Office* (EPO) e *Japan Patent Office* (JPO). A base de dados selecionada foi a USPTO por ser bastante representativa, considerando que reivindicações enviadas a outros países são frequentemente enviadas simultaneamente aos Estados Unidos, se mostrando uma base expressiva para o mercado tecnológico ao nível internacional (Bass & Kurgan, 2010). A coleta deste estudo estende-se às patentes referentes aos meses de janeiro a novembro do ano de 2019. O conjunto de dados para este período selecionado contêm 349.300 patentes, disponíveis no formato XML.

A terceira etapa é referente a indexação dos dados coletados utilizando a plataforma Apache Solr®. Os vários documentos XML de patentes são enviados para um sistema

gerenciador de banco de dados (SGBD) para posterior indexação. O SGBD utilizado neste trabalho foi o PostgreSQL® onde, é realizada uma varredura na tabela relacional armazenada no SGBD e, após a recuperação de cada linha, é enviada uma solicitação ao servidor Solr® para que ocorra a indexação. O documento de patentes importado contém a seguinte estrutura: identificador (*id*), título (*title*), data (*date*), ano (*year*), resumo (*abstract*), descrição (*description*) e tipo (*type*).

Na quarta etapa foi realizado o agrupamento dos dados (*clusters*) com o auxílio da ferramenta Carrot2®. O projeto Carrot2® permite organizar automaticamente coleções de documentos em categorias temáticas. Nesta pesquisa utilizamos o algoritmo Lingo®. Segundo (Osiński, Stefanowski, & Weiss, 2004), o Lingo® é um algoritmo utilizado para agrupamento, que extrai frases frequentes de documentos de entrada e, com base nas descrições, determina seu conteúdo através da redução da matriz termo-documento original utilizando *Singular Value Decomposition* (SVD). Os rótulos dos agrupamentos devem ser perceptíveis ao usuário para depois serem atribuídos aos documentos. O Lingo® possui 5 fases: a) pré-processamento; b) extração de frases frequentes; c) inserção de rótulos nos agrupamentos; d) descoberta do conteúdo dos agrupamentos e, a fase final e) formação dos agrupamentos.

Na quinta etapa, após a geração dos agrupamentos, utilizou-se a ferramenta de visualização *Foam Tree Visualization* disponível no Carrot2® para exibição dos resultados. Com isso foi possível observar como as patentes relacionadas à nanotecnologia se agrupam em tópicos na base de patentes.

Por fim, para a investigar a abrangência do tema na base de patentes foi verificada a relação entre os documentos presentes nesses agrupamentos com a classificação da *World Intellectual Property Organization* (WIPO). A WIPO classifica todos os pedidos de patentes publicados pela área tecnológica a qual pertencem. Essa classificação é denominada Classificação Internacional de Patentes (do inglês *International Patent Classification* - IPC) que codifica a aplicação e a função de determinada patente (WIPO, 2018). Dessa forma, pode-se investigar a quais seções se estendem os documentos pertencentes a cada agrupamento.

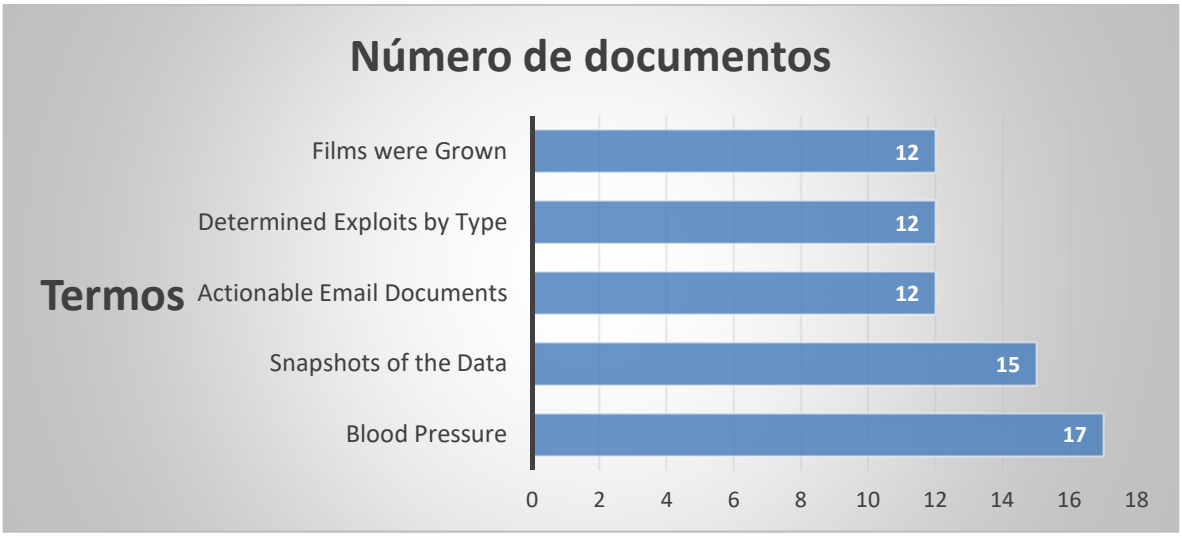
3. RESULTADOS E DISCUSSÃO

Com o objetivo de gerar agrupamentos de patentes com temas afins na área de nanotecnologia, apresentamos nesse estudo um processo de mineração de texto na base de patentes americanas provida pela USPTO referente ao ano de 2019. Com a indexação da base de dados concluída, foi realizada a consulta com a palavra-chave "*nanotechnology*" resultando em um total de 1.091 documentos de patentes.

Seguindo as etapas descritas na seção anterior, utilizou-se o software Carrot2® com as seguintes configurações: fonte de consulta foi o Solr®, algoritmo Lingo®, mapeamento do campo de índice utilizamos o *id* como campo de identificação, *description* para o campo de resumo e *title* para o campo de título. Na seção de consulta de pesquisa foi utilizado o termo de consulta "*nanotechnology*" para encontrar agrupamentos levando-se em conta o retorno dos 100 documentos (patentes) mais relevantes.

Com as configurações especificadas, o algoritmo Lingo® encontrou 49 agrupamentos com a configuração de "*Search query*" para 100 resultados, o método de fatoração padrão *Nonnegative Matrix factorization ED Factory*, ponderação do termo utilizando *tf-idf* e a retirada de *stopwords*, todos atributos padrão da ferramenta Carrot2®. Na Figura 2, temos os 5 agrupamentos com o maior número de documentos.

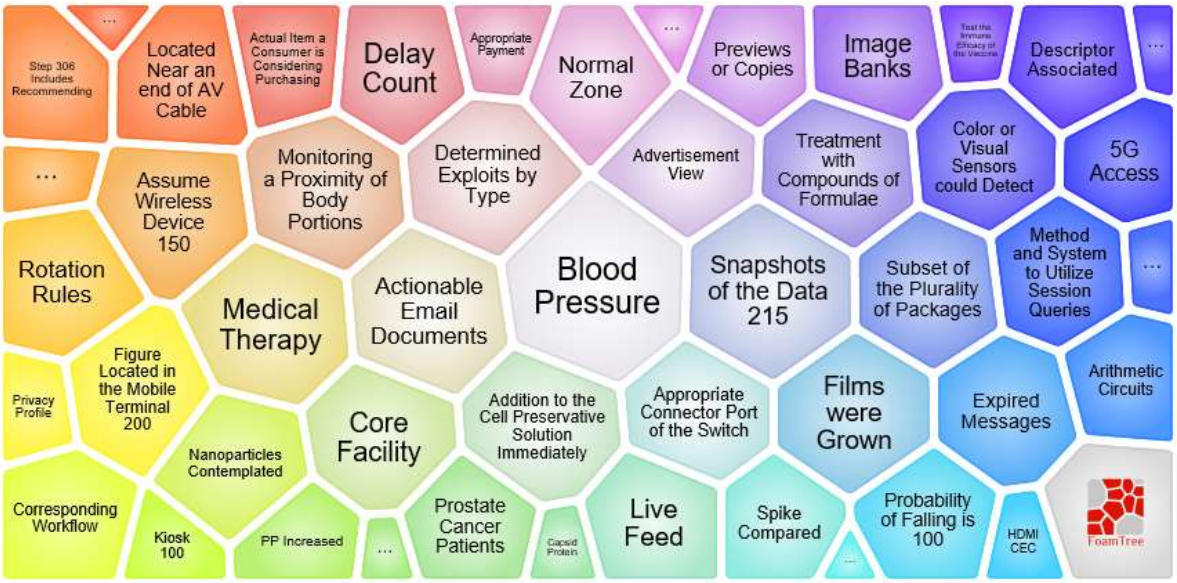
Figura 2 – Principais agrupamentos criados pelo Lingo.



Fonte: Os autores (2020).

Na Figura 3 têm-se os clusters resultantes da consulta sobre "*nanotechnolgy*" exibidos pela ferramenta de visualização *Foam Tree* disponível no Carrot2®. Os 49 agrupamentos mostram como se reúnem, por similaridade, as patentes na base de dados USPTO de janeiro a novembro de 2019.

Figura 3 - Agrupamentos de patentes para o termo nanotechnology.



Fonte: Os autores (2020).

A partir disso, foram selecionados os 5 primeiros agrupamentos para uma investigação mais detalhada verificando a relação com a seção e classificação IPC. Os agrupamentos selecionados foram '*Blood Pressure*' (17 documentos), '*Snapshots of the Data*' (15 documentos), '*Actionable Email Documents*' (12 documentos), '*Determined Exploits by Type'* (12 documentos) e '*Films were Grown'* (12 documentos), no total foram 68 documentos analisados.

O sistema de classificação do IPC divide as tecnologias em oito classes, de A a H, com cerca de 70.000 subdivisões (WIPO, 2018). Cada subdivisão possui um símbolo que consiste em algarismos arábicos e letras do alfabeto latino. A hierarquia consiste em: seção, classe, subclasse e grupo, que se subdivide em grupo principal e subgrupo. Seguindo essa classificação, entre os 68 documentos selecionados, foram identificadas quatro seções e treze classes.

As seções encontradas foram: 'A' referente a Necessidades Humanas, seção 'C' Química; Metalurgia, 'G' pertencente à seção Física e 'H' da seção Eletricidade. Na Tabela 1, podemos verificar que a maioria das pesquisas em nanotecnologia nestes 5 agrupamentos selecionados pertencem à seção Física.

Tabela 1 - Seções de patentes.

Seção	Descrição	Porcentagem
A	Necessidades humanas	1.47 %
C	Química; Metalurgia	19.12 %
G	Física	69.12 %
H	Eletricidade	10.29 %

Fonte: Os autores (2020).

Na Tabela 2, têm-se as classes encontradas nos 5 agrupamentos selecionados, por exemplo, a classe A61: pertence à seção A – Necessidades Humanas e a classe A61 – Ciência médica ou Veterinária; Higiene.

Tabela 2 - Classes de patentes.

Classe	Descrição
A61	CIÊNCIA MÉDICA OU VETERINÁRIA; HIGIENE
C07	QUÍMICA ORGÂNICA
C08	COMPOSTOS MACROMOLECULARES ORGÂNICOS; SUA PREPARAÇÃO OU SEU PROCESSAMENTO QUÍMICO; COMPOSIÇÕES BASEADAS NOS MESMOS
C12	BIOQUÍMICA; CERVEJA; ÁLCOOL; VINHO; VINAGRE; MICROBIOLOGIA; ENZIMOLOGIA; ENGENHARIA GENÉTICA OU DE MUTAÇÃO
C01	MEDIÇÃO; TESTE
G03	FOTOGRAFIA; CINEMATOGRAFIA; TÉCNICAS SEMELHANTES UTILIZANDO ONDAS OUTRAS QUE NÃO ONDAS ÓPTICAS; ELETROGRAFIA; HOLOGRAFIA
G08	SINALIZAÇÃO
G06	CÔMPUTO; CÁLCULO; CONTAGEM
G09	EDUCAÇÃO; CRIPTOGRAFIA; APRESENTAÇÃO VISUAL; ANÚNCIOS; LOGOTIPOS
G11	ARMAZENAMENTO DE INFORMAÇÕES

G16	TECNOLOGIA DE INFORMAÇÃO E COMUNICAÇÃO [ICT] ESPECIAL ADAPTADA PARA CAMPOS DE APLICAÇÃO ESPECÍFICOS
H01	ELEMENTOS ELÉTRICOS BÁSICOS
H04	TÉCNICA DE COMUNICAÇÃO ELÉTRICA

Fonte: Os autores (2020).

Na Tabela 3 são apresentadas as quantidades de classes encontradas nos 68 documentos conjuntamente com o percentual que representa. As classes com maior representatividade são referentes a seção Física, G01 – Medição e Teste e G06 - Cômputo; Cálculo; Contagem com 19,12% e 36,76%, respectivamente.

Tabela 3 - Contagem de classes.

Classe	Contagem de Classe	Porcentagem
A61	1	1.47 %
C07	4	5,88%
C08	2	2,94%
C12	7	10,29%
G01	13	19,12%
G03	1	1,47%
G06	25	36,76%
G08	4	5,88%
G09	2	2,94%
G11	1	1,47%
G16	1	1,47%
H01	2	2,94%
H04	5	7,35%

Fonte: Os autores.

O código IPC não era campo indexado do Solr® e foi obtido através da consulta do título do documento indexado nos sites da USPTO e Google Patents®. Após a verificação do título, foi analisado no campo descrição, o número do registro da patente. Com isso, os dados foram extraídos para uma planilha para se chegar aos resultados apresentados neste artigo.

As tabelas 1 e 3 trazem uma visão geral dos documentos pertencentes aos 5 principais agrupamentos. Por sua vez, a tabela 4 apresenta uma distribuição segmentada pelos agrupamentos encontrados. Pode-se observar a prevalência da seção G referente à Física, no agrupamento '*Actionable Email Documents'* o qual, averiguando os documentos que o compõe, percebe-se que o agrupamento trata de patentes relativas ao tema de fluxo de informações. Já o agrupamento '*Determined Exploits by Type*' é o único que abrange as quatro seções.

Tabela 4 - Classes por agrupamento de patentes.

Agrupamento	**Seção**	**Porcentagem**
Blood Pressure	A	-
	C	29.41 %
	G	64.70 %
	H	5.88 %
Snapshots of the Data	A	-
	C	-
	G	86.67 %
	H	13.33 %
Actionable Email Documents	A	-
	C	-
	G	91.67 %
	H	8.33 %
Determined Exploits by Type	A	8.33 %
	C	33.33 %
	G	50 %
	H	8.33 %
Films were Grown	A	-
	C	33.33 %
	G	50 %
	H	16.67 %

Fonte: Os autores (2020).

O domínio da nanotecnologia abrange os campos da ciência, engenharia e tecnologia. Segundo o NNI, estudos da nanotecnologia se concentram em imagem, medição, modelagem e manipulação de matéria em nanoescala. De forma que as classificações dos resultados apresentados pelos 5 principais agrupamentos corroboram com essa informação. Sendo as classes referentes à medição, teste, cômputo, cálculo e contagem as que se encontram majoritariamente dentre as parentes presentes nos agrupamentos.

Entretanto, por outro lado, observando os rótulos atribuídos pelo algoritmo Lingo®, na Figura 3, fica evidente o atual interesse da aplicação da nanotecnologia para a área da saúde. É possível notar 7 agrupamentos associados à temática da saúde; '*Blood preasure*', '*Prostate cancer patients*', '*Medical therapy*', '*Treatment with compounds of formulae*', '*Addition to the cell preservative solution immediately*', '*Monitoring a proximity of body portions*' e '*Capsid protein*'. Assim, como apresenta o estudo bibliográfico, nos últimos anos a nanotecnologia têm sido aplicada à saúde humana com resultados promissores, especialmente no campo do tratamento de câncer (Porter et al., 2019; Bayda et al., 2020).

Tal resultado permite inferir que das patentes requeridas no ano de 2019 na base USPTO, uma parcela significativa devem se referir ao desenvolvimento de processos de medição voltados para a aplicação na área da saúde. O estudo de (Stopar et al. 2016) investigando as disciplinas envolvidas nas pesquisas de nanociência e nanotecnologia, usando a base bibliográfica (*Web of Science*), similarmente, identificou como sendo os dois grupos mais relevantes 'bio' e 'nano', ou seja, estudos biológicos e médicos relacionados à física, química e materiais. Tais informações apoiam os resultados atingidos neste trabalho. Em Bayda et al. (2020) percebe-se a proximidade dos conceitos nanotecnologia e nanociência o processo de desenvolvimento de materiais (modelagem, cálculos e medições) e aplicações são desenvolvimentos paralelos. Justificando o achado, de forma que seja por meio da base de patentes ou na literatura, existe o desenvolvimento relativo à física e química concomitante com a identificação das aplicações, no momento atual, referentes em especial para a área da saúde e energia.

4. CONCLUSÃO

Com o objetivo de realizar uma análise exploratória de como o termo 'nanotecnologia' se organiza na base de patentes, foi realizado um processo de mineração de texto na base americana USPTO referente ao ano de 2019. A consulta a partir do termo "*nanotechnology*" resultou em 1.091 documentos de patentes. Os 5 principais agrupamentos encontrados pelo algoritmo Lingo®, a partir das frases frequentes dos documentos indexados foram: '*Blood*

Pressure', '*Snapshots of the Data'*, '*Actionable Email Documents'* , '*Determined Exploits by Type'* e '*Films were Grown'*.

A fim de investigar a abrangência desses agrupamentos na base de patentes, uma verificação referente a classificação IPC identificou 4 seções, sendo: 'A' - Necessidades Humanas, 'C' - Química; Metalurgia, 'G' – Física; e 'H' - Eletricidade. A seção G (Física) se mostrou a mais significativa dentre os 5 principais agrupamentos, ainda que a visualização dos agrupamentos apontem para a relação da nanotecnologia com a área da saúde. Deste modo, pode-se deduzir que processos de medição voltados à área da saúde foram o foco das patentes em 2019. Sendo assim, seguindo as etapas descritas na metodologia deste artigo, consideramos que a técnica utilizada foi eficiente.

Das limitações que a metodologia implica, as etapas 2 e 3 envolvendo a seleção e a indexação do conjunto de dados, respectivamente, foram as etapas mais árduas da pesquisa. Além do esforço computacional, é fundamental o correto pré-processamento dos dados visando obter consistência e credibilidade para a pesquisa. Como visto a nanotecnologia se estende a muitos campos da ciência, engenharia e tecnologia, tornando desafiador seu mapeamento e organização. Tal que é difícil acessar todas as patentes relacionadas à nanotecnologia, pois existem termos como "*quantum dot*", "*graphene*", "*fullerene*" entre outros, que podem não ser abarcados por buscas com o termo "*nanotechnology*", embora sejam termos de importância para o campo da nanotecnologia.

Todavia, o estudo tem sua contribuição, pois diferentemente de estudos que limitam suas investigações a partir do enquadramento da nanotecnologia pela classificação IPC, seção B82, este estudo se comprometeu em explorar a extensão do termo nanotecnologia na base de patentes.

Estudos futuros podem se comprometer a investigar todas as classes para os 49 clusters encontrados, seguindo o código IPC. Ou ainda, estima-se que a utilização de termos mais específicos possam retornar resultados mais relevantes. Considerando a proximidade dos conceitos nanociência e nanotecnologia, talvez uma análise com a consulta pelo termo "nanomaterial" traga resultados interessantes na base de patentes por meio do agrupamento de áreas com foco na aplicação e desenvolvimento desses nanocompostos. Por fim, é realizada a recomendação para estudos que considerem a evolução temporal desses agrupamentos.

5. AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

REFERÊNCIAS

Abbas, A., Zhang, L., & Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. *World Patent Information*, *37*, 3–13. https://doi.org/10.1016/j.wpi.2013.12.006

Afantitis, A. (2020). Nanoinformatics: Artificial Intelligence and Nanotechnology in the New Decade. *Combinatorial Chemistry & High Throughput Screening*, *23*(1), 4–5.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. *Undefined*, *13*. Retrieved from http://en.wikipedia.org/wiki/Statistics

Antons, D., Grünwald, E., Cichy, P., & Salge, T. O. (2020). The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *R and D Management*, *50*(3), 329–351. https://doi.org/10.1111/radm.12408

Barnard, A. S., Motevalli, B., Parker, A. J., Fischer, J. M., Feigl, C. A., & Opletal, G. (2019, November 7). Nanoinformatics, and the big challenges for the science of small things. *Nanoscale*, Vol. 11, pp. 19190–19201. https://doi.org/10.1039/c9nr05912a

Bass, S. D., & Kurgan, L. A. (2010). Discovery of factors influencing patent value based on machine learning in patents in the field of nanotechnology. *Scientometrics*, *82*(2), 217–241. https://doi.org/10.1007/s11192-009-0008-z

Bayda, S., Adeel, M., Tuccinardi, T., Cordani, M., & Rizzolio, F. (2020). The history of nanoscience and nanotechnology: From chemical-physical applications to nanomedicine. *Molecules*, Vol. 25. https://doi.org/10.3390/molecules25010112

Charreau, H., Cavallo, E., & Foresti, M. L. (2020, June 1). Patents involving nanocellulose: Analysis of their evolution since 2010. *Carbohydrate Polymers*, Vol. 237, p. 116039.

Ernst, H. (2003). Patent information for strategic technology management. *World Patent Information*, *25*(3), 233–242. https://doi.org/10.1016/S0172-2190(03)00077-2

Fayyad, U., & Uthurusamy, R. (1996). Data Mining and Knowledge Discovery in Databases. *Communications of the ACM*, *39*(11), 24–26. https://doi.org/10.1145/240455.240463

Gao, J. ping, Ding, K., Teng, L., & Pang, J. (2012). Hybrid documents co-citation analysis: Making sense of the interaction between science and technology in technology diffusion. *Scientometrics*, *93*(2), 459–471. https://doi.org/10.1007/s11192-012-0691-z

INPI. (2013). *Instituto Nacional da Propriedade Industrial (Brasil)*. Retrieved from https://www.gov.br/inpi/pt-

br/composicao/arquivos/03_cartilhapatentes_21_01_2014_0.pdf

Kah, M., Tufenkji, N., & White, J. C. (2019, June 1). Nano-enabled strategies to enhance crop nutrition and protection. *Nature Nanotechnology*, Vol. 14, pp. 532–540.

Li, X., Fan, M., Zhou, Y., Fu, J., Yuan, F., & Huang, L. (2020). Monitoring and forecasting the development trends of nanogenerator technology using citation analysis and text mining. *Nano Energy*, *71*, 104636. https://doi.org/10.1016/j.nanoen.2020.104636

Ma, J., Abrams, N. F., Porter, A. L., Zhu, D., & Farrell, D. (2019). Identifying translational indicators and technology opportunities for nanomedical research using tech mining: The case of gold nanostructures. *Technological Forecasting and Social Change*, *146*, 767–775.

Madani, F., & Weber, C. (2016). The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis. *World Patent Information*, *46*, 32–48.

NNI, N. N. I. (2012). Nanotechnology Knowledge Infrastructure: Enabling National Leadership in Sustainable Design. *NSTC Committee on Technology—Subcommittee of Nanoscale Science, Engineering and Technology*.

Osiński, S., Stefanowski, J., & Weiss, D. (2004). Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. In *Intelligent Information Processing and Web Mining* (pp. 359–368). https://doi.org/10.1007/978-3-540-39985-8_37

Panneerselvam, S., & Choi, S. (2014). Nanoinformatics: Emerging databases and available tools. *International Journal of Molecular Sciences*, *15*(5), 7158–7182.

Porter, A. L., Garner, J., Newman, N. C., Carley, S. F., Youtie, J., Kwon, S., & Li, Y. (2019). National nanotechnology research prominence. *Technology Analysis and Strategic Management*, *31*(1), 25–39. https://doi.org/10.1080/09537325.2018.1480013

Royal Society. (2004). Nanoscience and Nanotechnologies: Opportunities and Uncertainties. In *London: The Royal Society*.

Shalaby, W., & Zadrozny, W. (2019). Patent retrieval: a literature review. *Knowledge and Information Systems*, *22*(3), 1–30. https://doi.org/10.1007/s10115-018-1322-7

Singh, V., Chakraborty, K., & Vincent, L. (2016). Patent Database: Their Importance in Prior Art Documentation and Patent Search. Retrieved June 20, 2020, from Journal of Intellectual Property Rights website: http://nopr.niscair.res.in/handle/123456789/34016

Stopar, K., Drobne, D., Eler, K., & Bartol, T. (2016). Citation analysis and mapping of nanoscience and nanotechnology: identifying the scope and interdisciplinarity of research. *Scientometrics*, *106*(2), 563–581. https://doi.org/10.1007/s11192-015-1797-x

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, *43*(5), 1216–1247.

WIPO. (2018). International Patent Classification (IPC). *World Intellectual Property Organization*, *1*, 1. https://doi.org/10.1007/978-1-4614-8351-9_16