A method toward co-word network analysis in patent databases

Um método voltado à análise de rede de termos em bases de patentes

ABSTRACT
Knowledge discovery through word co-occurrence networks provides resources for experts to access a certain knowledge domain and take proper decisions in order to maximize and define investments. Usually co-word analysis applies to scientific papers. However, the unstructured part of patent documents are promising for this kind of analysis. Therefore, this work proposes a six-step method using text mining and network analysis techniques. The method provides steps that can assist the investigation and understanding of technology networks that are formed based on a topic of interest. In that way, was carried out a study focused on analyzing the interrelationships of terms from patent documents in the domain of nanotechnology. The theme was picked considering its relevance and scope of applications, as well as the benefits for society. Thus, three main communities were identified, two of which are related to health and one to energy. The communities generated, as well as the refinement of the dictionary, tend to facilitate the understanding of a knowledge area. Finally, it is emphasized that patent analysis has the potential to improve or define new products and processes impacting the competitiveness of technology-based organizations.

Key words: social network analysis; patent analysis; co-word network; nanotechnology; text mining

RESUMO
A análise de rede de coocorrência de termos permite a descoberta de conhecimento promovendo subsídios para que especialistas possam avaliar adequadamente determinada área e determinar caminhos para maximizar e definir investimentos. De modo geral, a análise de coocorrência de termos é usualmente aplicada a bases de publicações científicas. Todavia, o conteúdo não estruturado presente em patentes se mostra adequado e promissor para este tipo de análise. Neste sentido, este trabalho propõe um método composto por seis etapas utilizando-se de técnicas de mineração de texto e de análise de redes. O método fornece passos que podem auxiliar na investigação e entendimento de redes de tecnologia que se formam a partir de uma temática de interesse. Dessa forma, um estudo visando analisar as inter-relações de termos a partir de documentos de patente para o domínio da nanotecnologia foi efetuado. Esta temática foi definida levando-se em conta sua relevância e abrangência de aplicações, assim como os benefícios para a sociedade. Como resultado, três principais comunidades foram identificadas, sendo duas delas relativas à área da saúde e uma na área de energia. As comunidades geradas, bem como o refinamento do dicionário, tendem a facilitar o entendimento de uma área. Por fim, ressalta-se que a análise de patentes tem potencial para melhorar ou definir novos produtos e processos impactando na competitividade de organizações de base tecnológica.

Palavras-chave: análise de redes sociais; análise de patentes; rede de coocorrência de termos; nanotecnologia; mineração de texto

1. INTRODUÇÃO E CONTEXTUALIZAÇÃO

Uma forma de descoberta de conhecimento que tem atraído atenção dos pesquisadores é a análise das interconexões presentes nas informações disponíveis na *web*. A análise de redes proporciona um modo sistemático de revelar estruturas e conhecimentos ocultos que seriam difíceis de observar por outros meios (Newman, 2003). Atuando na representação de informações e na análise visual, permite uma melhor representação dos conjuntos de dados visando revelar padrões ocultos nesses dados (Liu et al., 2018). Ainda, para Ernst (2003), na busca de oportunidades tecnológicas, a visualização da informação é uma tarefa crucial para apresentar resultados de fácil compreensão. Uma rede permite a visualização da relação entre elementos ou atores a partir de sua representação por meio de nós (vértices) e suas conexões (arestas). Possibilita assim, melhores decisões, bem com a identificação de oportunidades mediante a análise do posicionamento dos nós e suas conexões na rede.

A análise de redes surge do desdobramento da Teoria de Grafos como principal linguagem matemática para mensurar as propriedades das redes (Newman, 2003). Entretanto, a análise de redes faz uso também das teorias e métodos da física estatística, ciência da computação, estatística e sociologia (Barabási, 2013). Isso em razão de entender como emergem e evoluem redes reais, ou seja, as redes que se formam na natureza, tecnologia e sociedade (Newman, 2003). Dessa forma, a ciência de redes tem como premissa, que apesar das diferenças aparentes entre essas redes, elas satisfazem um conjunto fundamental de leis e mecanismos que podem ser interpretados a partir de um conjunto unificado de ferramentas e princípios (Barabási, 2013).

A retomada de interesse na análise de redes, enfatizando as características presentes nas redes reais, surge no final dos anos 1990, e pode ser atribuída aos trabalhos de Duncan J. Watts e Steven Strongatz (1998) e Albert-László Barabási e Réka Albert (1999), nos quais apresentam os princípios de *Small-World* (Watts & Strogatz, 1998) e Escala-livre (Barabási & Albert, 1999). Respectivamente, mostrando que nas redes reais, diferentemente da teoria tradicional (Erdös & Rényi, 1959), a distância média entre os nós permanece pequena, mesmo quando a rede se torna muito grande (Watts & Strogatz, 1998). Assim como, a maioria dos nós apresenta um número de conexões muito baixo, ao mesmo tempo que, por outro lado, alguns nós são altamente conectados (*hubs*) (Barabási & Albert, 1999). Essas características podem ser encontradas especialmente ao se analisar redes sociais. A partir de 2000, Girvan e Newman, (2002) exploram a ciência de redes sob o aspecto social, contribuindo para a difusão da técnica de análise de redes. Por essa razão, muitas pesquisas que fazem uso de análise de redes a nomeiam como análise de redes sociais (SNA - *social network analysis*).

A análise de redes reais, ou redes complexas, é uma área relativamente recente. Pode-se indicar como marco o ano de 2005 quando o Conselho Nacional de Pesquisa dos EUA definiu a ciência de redes como um novo campo de pesquisa básica (Molontay & Nagy, 2019). A área percorreu um longo caminho para estabelecer uma base em comum, chegar a um acordo sobre definições e conciliar abordagens adotadas por campos tão díspares quanto ciências sociais, física, biologia, ciência da computação e matemática aplicada. Contudo, apesar das dificuldades inerentes a campos interdisciplinares, a área de ciência de redes apresenta avanço e consolidação (Vespignani, 2018; Molontay & Nagy, 2019). Sendo promissora, as pesquisas atuais se beneficiam da disponibilidade do poder computacional e de grandes bases de dados. Promovem ainda um aumento no potencial de análise indo desde a dinâmica de nós individuais até as

propriedades emergentes de redes macroscópicas (Vespignani, 2018) nas mais variadas temáticas que, no contexto deste artigo se refere a nanotecnologia.

Como temática de investigação, a nanotecnologia se mostra atrativa e relevante, tanto para a ciência quanto para a tecnologia. Tendo em vista a abrangência das aplicações e possíveis benefícios para a sociedade, inerente à sua definição, pela habilidade de manipular e controlar individualmente átomos e moléculas (Bayda, Adeel, Tuccinardi, Cordani, & Rizzolio, 2020). Uma vez que, tudo é composto por átomos, pode-se especular sobre o potencial da área, justificando o grande interesse de pesquisa e investimento, tanto por pesquisadores da área acadêmica, como por organizações privadas e centros de Pesquisa e Desenvolvimento (P&D), assim como por Governos (Jiang, Gao, Chen, & Roco, 2015).

O campo da nanotecnologia teve seu início em 1959, com a publicação de Richard Feynman *"There's plenty of room at the bottom"*, descrevendo as possibilidades que poderiam ocorrer se os cientistas pudessem aprender a controlar átomos e moléculas (Toumey, 2009). Em 1989, o periódico *Nanotechnology* foi fundado, consolidando o termo na comunidade científica (Toumey, 2009). Nesse mesmo ano, Don Eigler realizou a primeira manipulação atômica na IBM®, considerado um marco importante para a evolução da pesquisa na área (Fanfair, Desai, & Kelty, 2007). Em 2000, a nanotecnologia foi proclamada uma prioridade para as pesquisas nos Estados Unidos, fundando a *National Nanotechnology Initiative* (NNI) (Porter et al., 2019). O campo seguiu evoluindo no século XXI abarcando pesquisas e aplicações em diversas áreas, da física à medicina (Kagan et al., 2016), acumulando assim uma grande quantidade de publicações científicas, relatórios técnicos e patentes.

Em especial, a base de dados de patentes é de grande valia para investigação visando a descoberta de conhecimento. Documentos de patentes representam uma invenção, um produto ou um processo, que fornece uma nova solução tecnológica para resolver determinado problema (Madani & Weber, 2016). Mais que isso, documentos de patentes contém resultados tecnológicos raramente replicados em outras publicações, em razão de trazer importantes resultados de pesquisa sobre algo novo, sem anterioridade (INPI, 2013). Desta forma, a análise de patentes desempenha um papel fundamental para fins de planejamento estratégico, incluindo a descoberta de tecnologias promissoras, avaliação de avanços tecnológicos e novas tendências (Abbas, Zhang, & Khan, 2014).

A análise de patentes 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, Chakraborty, & Vincent, 2016). A análise de patentes possibilita a extração de informações a partir de dados estruturados e não estruturados (Tseng, Lin, & Lin, 2007). Para a análise da parte não estruturada das patentes, ou seja, o texto, técnicas de mineração de texto são usadas para extrair informações. Enquanto as técnicas de visualização se ocupam da representação visual das informações da patente auxiliando especialistas nas tomadas de decisão.

O campo da nanotecnologia apresenta uma relação próxima tanto com avanço na ciência quanto com o desenvolvimento tecnológico (Gao, Ding, Teng, & Pang, 2012). Nessa perspectiva o uso da base de patentes é presente como fonte de dados para análise de redes nos estudos que investigam o domínio da nanotecnologia. Entretanto, os estudos ocorrem principalmente na análise da rede de colaboração entre atores (Chang, 2018; Jiang et al., 2015; Ozcan & Islam, 2014) e na análise do fluxo de conhecimento por meio da citação entre patentes, uma vez que cada patente traz referências do que já existe (Wang & Guan, 2011; Zingg & Fischer, 2018, 2019). Dessa forma, por mais que a técnica de análise de redes e a base de

patentes venham sendo amplamente utilizadas para investigar o campo da nanotecnologia, existe ainda a necessidade de se explorar a rede de termos do domínio de conhecimento da nanotecnologia.

As redes de coocorrência de termos, ou *co-word analysis*, são importantes para a compreensão dos componentes do conhecimento e da estrutura do conhecimento de um campo científico ou técnico, examinando as conexões entre os termos na base de dados (Li, Zhang, & Hong, 2020; Su & Lee, 2010). Mais ainda, a visualização das relações entre palavras na forma de rede é destacada por fornecer entendimento intuitivo (Katsurai & Ono, 2019). E embora sejam aplicadas usualmente a bases de publicações científicas (Muñoz-Écija, Vargas-Quesada, & Chinchilla-Rodríguez, 2017; Radhakrishnan, Erbis, Isaacs, & Kamarthi, 2017), a parte não estruturada das patentes fornece uma base adequada e promissora para este tipo de análise. Nesse sentido, pode-se explorar a descoberta de conhecimento em texto a partir da análise dos relacionamentos estabelecidos entre os elementos textuais dos documentos de patentes.

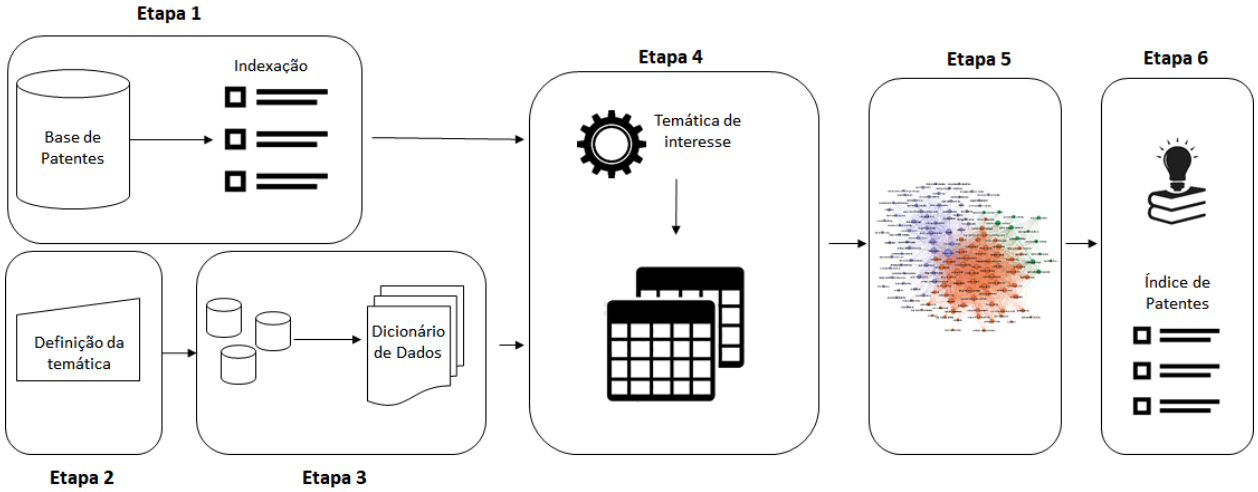
Diante disso, este estudo propõe um método para investigar a inter-relação de termos nos documentos de patente para o domínio da nanotecnologia. Para tanto, o estudo objetiva processar, via técnica de mineração de texto, os documentos associados ao tópico 'nanotecnologia' a partir da base de patentes americanas USPTO, referentes aos anos de 2019 e 2020. Bem como, investigar as inter-relações por meio da análise da rede de coocorrência de termos.

Para a análise da rede de coocorrência de termos utilizou-se medidas de centralidade de grau, centralidade de proximidade e centralidade de intermediação (Freeman, 1977), medidas essenciais para investigar e descrever o posicionamento de um nó na rede. No presente estudo, foi também explorado o conceito de modularidade, que de maneira visual possibilita a identificação de subdivisões nas quais os nós se organizam dentro de uma rede, igualmente conhecido por comunidades. Proporcionam uma visão ampla do conhecimento envolvido no domínio da nanotecnologia, favorecendo novos *insights* para a área. O artigo prossegue apresentando o processo metodológico aplicado, em sequência, os resultados e discussões. Finaliza com a conclusão e proposições para estudos futuros.

2. MÉTODO PROPOSTO

Com a finalidade de atender a lacuna identificada por esta pesquisa sugerimos um método. Este é composto por seis etapas principais combinando técnicas de mineração de texto, análise de redes e visualização da informação, que juntas possibilitam a análise da interconexão de termos a partir de uma temática de interesse visando a descoberta de conhecimento. O encadeamento das etapas é descrito na Figura 1.

Figura 1 - Etapas do método proposto baseados em Mineração de Texto e Análise de Redes



Fonte: Os autores (2020)

A primeira etapa é responsável pela seleção, coleta e indexação dos documentos de patentes. A base de patentes selecionada foi a americana, *United States Patent and Trademark Office* (USPTO). Tal escolha se dá por sua representatividade, considerando que reivindicações enviadas a outros países são frequentemente submetidas 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 anos de 2019 e 2020 (até a primeira semana de junho de 2020). O conjunto de dados para este período selecionado contêm 564.959 patentes, disponíveis no formato XML. Em seguida, foi realizada a indexação das patentes coletadas utilizando a plataforma Apache Solr®. Cada documento de patente indexado contém a seguinte estrutura: identificador (*id*), título (*title*), data (*date*), ano (*year*), resumo da patente (*abstract*) e descrição da patente (*description*). O resultado dessa etapa é a base de patentes indexada possibilitando a consulta por palavras-chave.

Na segunda etapa, a qual ocorre independente da primeira, temos a definição de uma temática de investigação. No caso, este estudo refere-se ao domínio da Nanotecnologia pela motivação apresentada na seção anterior. Mais especificamente, para a elaboração do cenário foi utilizado o termo *nanotecnolog**, em que o "*" indica todas as variações do termo.

Para a terceira etapa determina-se o dicionário dos termos de interesse, ou seja, uma lista com termos relevantes para determinado domínio de interesse. Para a pesquisa, a coleta dos termos foi realizada manualmente através da busca por artigos contendo o termo nanotecnologia, entre 2018 e 2019, nas bases *Web of Science* e *Google Scholar*. Durante a pesquisa, foram extraídas as palavras-chave desses artigos, bem como dicionários de termos utilizados na área de nanotecnologia. Após a composição inicial do dicionário foi realizada uma triagem a fim de eliminar duplicidades e termos genéricos que pudessem ser designados à várias áreas, totalizando 257 termos (Apêndice I).

A quarta etapa se ocupa do preenchimento da base de dados que suporta a análise de redes. Para tal, foi elaborado um modelo de dados composto de três tabelas. A primeira tabela permite armazenar os termos que constam no dicionário contendo um identificador e uma descrição (nome do termo). Associada a esta tabela existe a tabela de frequências sendo que cada termo (relacionado pelo identificador a tabela de termos) possui a quantidade de patentes

em que este é mencionado. Por fim, existe uma tabela que permite armazenar as coocorrências de dois termos quaisquer. Partindo de um termo de origem, são realizadas *n* consultas com os termos subsequentes (destino) na base de patentes indicando a quantidade de documentos que mencionam os dois termos conjuntamente, ou seja, a coocorrência.

A etapa cinco visa gerar e analisar a rede de coocorrência de termos. A análise da rede é constituída em dois passos: o contexto da temática e a rede de termos agrupados (comunidades) com base nos termos do dicionário que envolvem o contexto da temática. Aqui foi utilizada a ferramenta Gephi® para gerar a visualização e análise das medidas. A visualização permite a observação das relações dos termos sendo que quanto maior o número de documentos nos quais os termos coocorrem, mais forte será a relação entre estes termos (Grames, Stillman, Tingley, & Elphick, 2019). O pretexto dessa análise encontra-se no fato de quando termos ocorrem com certa frequência, próximos uns aos outros, estes têm uma relação mais forte do que termos que ocorrem com menos frequência. Dessa forma, os termos que coocorrem tendem a expressar a existência de temáticas recorrentes, assuntos centrais e conceitos que constituem e estruturam a área de estudo investigada (Li, Zhang, & Hong, 2020). Por sua vez, as medidas de centralidade de grau (número de conexões de um nó) e, a centralidade de intermediação (*betweenness*) e de proximidade (*closeness*), mensuram o posicionamento dos nós na rede. A centralidade de intermediação avalia o quanto determinado nó encontra-se no caminho para se atingir outros nós. Por outro lado, a centralidade de proximidade analisa quão central determinado nó é em relação a outros nós.

Prosseguindo, ainda na etapa cinco, para a determinação das comunidades utilizou-se o algoritmo de modularidade disponível na ferramenta Gephi®. A formação das comunidades depende da comparação da densidade de conexões dentro de um grupo de nós e deste grupo em relação ao restante da rede (Yang, Liu, & Liu, 2010). A investigação das comunidades permite uma compreensão complementar da rede, uma vez que, as comunidades podem ser vistas como meta-nodos dentro da rede (Newman & Girvan, 2004). Sendo que internamente uma comunidade pode apresentar um comportamento diferente quando comparado ao da rede como um todo. A observação das comunidades pode ainda contribuir para a identificação de conexões ausentes ou ainda conexões falsas, errôneas, dentro da rede.

Por fim, a etapa seis diz respeito à análise da base de patentes considerando os principais termos de cada comunidade identificada. Utilizando tais termos uma consulta é realizada com o intuito de recuperar os documentos mais aderentes/similares. A partir do resultado, os termos que já constam no dicionário são destacados no texto ficando, sob a responsabilidade do especialista/analista, a identificação de novos termos relacionados ao contexto da pesquisa que possam ser adicionados ao dicionário. Tal abordagem, objetiva o aprimoramento do dicionário com o intuito de promover um entendimento mais amplo sobre a temática de interesse, e neste aspecto, possibilitar aos especialistas da área obter novos *insights* que possam contribuir em áreas como a gestão da inovação, bem como auxiliar na tomada de decisão.

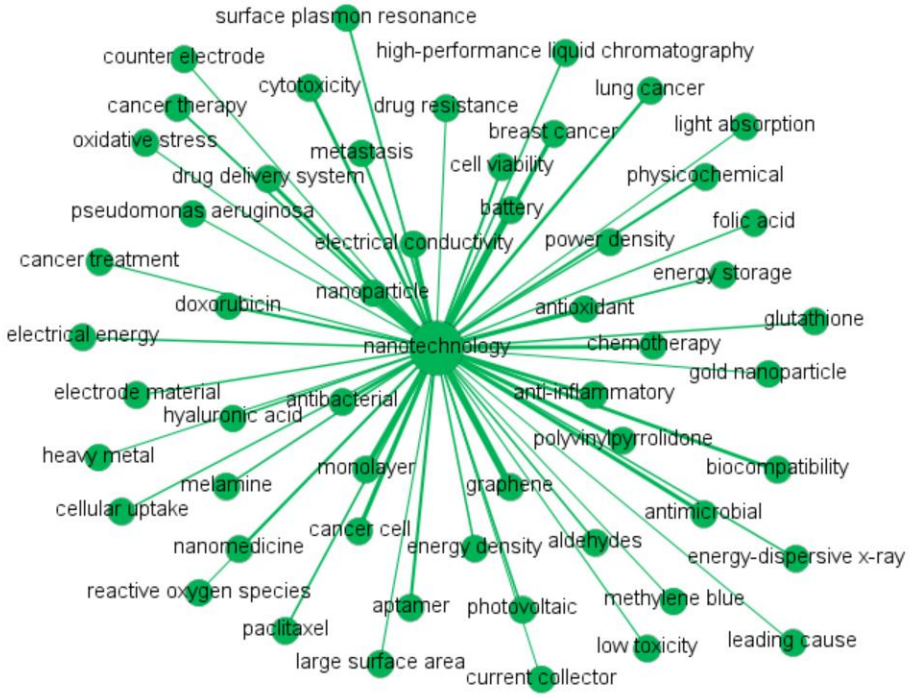
3. APLICAÇÃO DO MÉTODO E ANÁLISE DE RESULTADOS

Atendendo o propósito de analisar a inter-relação de termos nos documentos de patente para o domínio da nanotecnologia. A partir do conjunto de termos disponibilizados no dicionário foram elaboradas duas avaliações com base na Análise de Redes, sendo uma voltada à apresentação dos termos mais relacionados ao termo Nanotecnologia e outra com foco na

apresentação de agrupamentos de termos (comunidades) presentes no contexto de Nanotecnologia.

A Figura 2 apresenta uma rede com os termos mais associados à *nanotechnology*, com destaque para *nanoparticle*, *battery*, *graphene*, *monolayer e antioxidant.* Cada um dos termos é mencionado conjuntamente em 433, 280, 267, 210 e 172 patentes, respectivamente.

Figura 2 - Rede de termos relacionados à *nanotechnology*



Considerando cada um dos nós presentes no grafo pode-se analisar as associações mais relevantes visando compreender o contexto de cada um dos termos. O Quadro 1 apresenta os 5 (cinco) termos mencionados anteriormente e os 3 (três) termos mais importantes relacionados para cada um. O peso é determinado pela quantidade de patentes em que o termo e o seu relacionamento aparecem conjuntamente.

Quadro 1 - Informações gerais sobre as comunidades geradas

Termo Origem	Termo Relacionado	Peso
nanoparticle	*cancer cell*	114
	breast cancer	108
	monolayer	91
battery	*graphene*	64
	nanoparticle	53
	electrical energy	51
graphene	*nanoparticle*	82
	electrical conductivity	81
	battery	64
monolayer	*nanoparticle*	91
	graphene	61
	cancer cell	33
antioxidant	*nanoparticle*	61
	anti-inflammatory	54
	polyvinylpyrrolidone	53

A partir desse contexto inicial elaborou-se uma segunda rede contendo todos os termos presentes em patentes vinculadas à *nanotechnology*, ou seja, uma rede formada por termos que coocorrem no contexto de nanotecnologia. Por exemplo, para gerar determinada conexão entre dois termos quaisquer, "Termo 1" e "Termo 2", a busca na base de patentes agrega sempre o termo Nanotechnolog* ("Termo 1" AND "Termo 2" AND "Nanotechnolog*"). O retorno será a quantidade de patentes que satisfazem a expressão de busca que representará o peso da aresta entre "Termo 1" e "Termo 2".

Para permitir uma análise mais detalhada foram calculadas as medidas de centralidade de grau, intermediação e proximidade, assim como a aplicação do algoritmo de modularidade com o intuito de visualmente apresentar as comunidades geradas, que são vistas como agrupamentos. As informações gerais sobre as comunidades produzidas são apresentadas no Quadro 2.

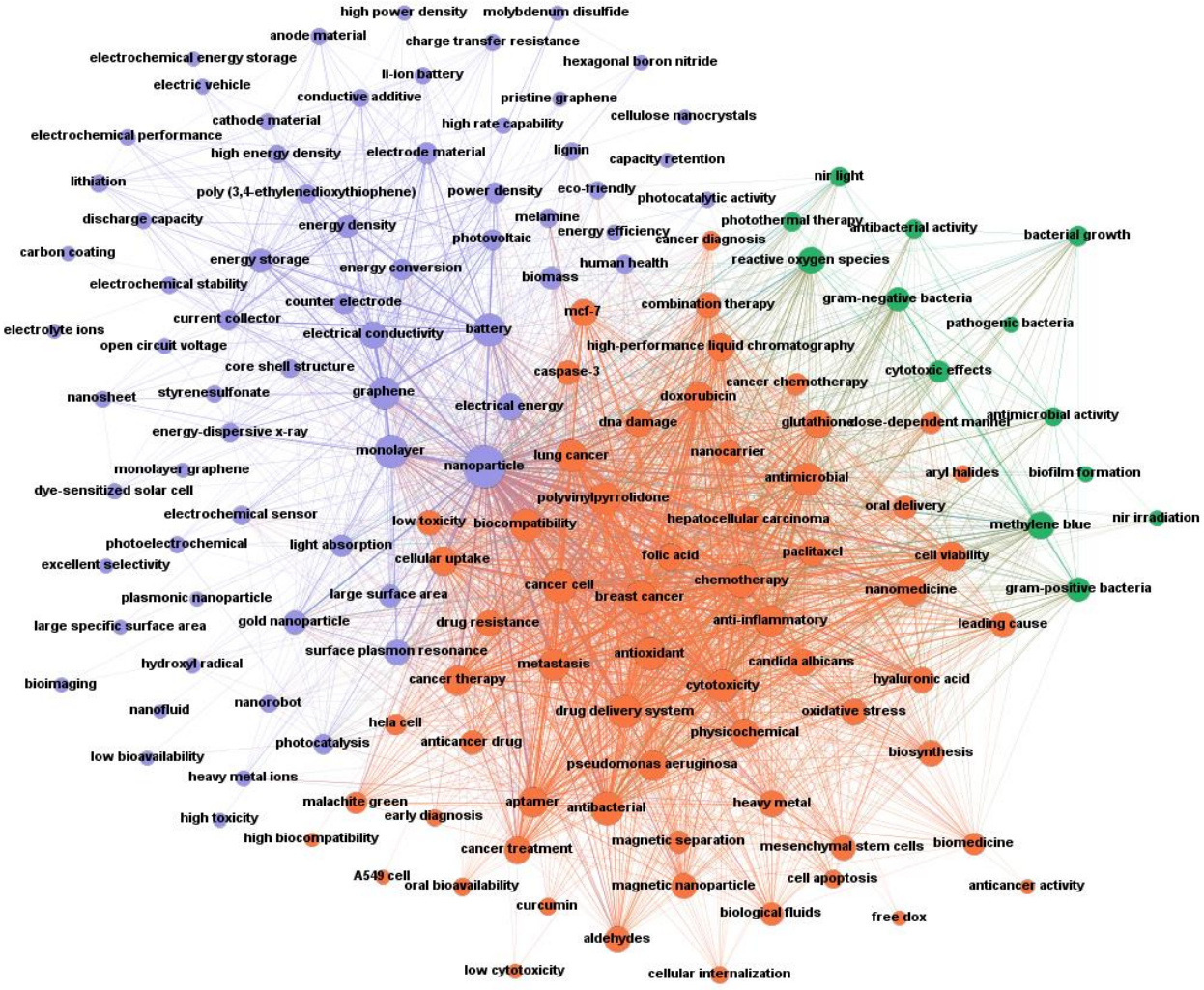
Quadro 2 - Informações gerais sobre as comunidades geradas

Identificador da Comunidade	Quantidade de nodos/termos	Percentual em relação à rede
0 (roxo)	67	46,53%
1 (laranja)	64	44,44%
2 (verde)	13	9,03%

Com relação à estrutura da rede, a Figura 3 exibe às três comunidades nas cores roxo (Comunidade 0), laranja (Comunidade 1) e verde (Comunidade 2). Para permitir uma melhor visualização somente termos com uma coocorrência (número de vezes que dois termos aparecem conjuntamente em uma patente) maiores ou iguais a 5 foram considerados. Este limiar foi definido testando valores até 10. Valores inferiores a 5 promovem uma rede com muitos

nodos, enquanto valores maiores reduzem a expressividade das comunidades identificadas. Analisando as principais estatísticas a rede possui um grau médio de 34,056, com diâmetro de 4 e densidade de 0,238. Os valores de diâmetro e densidade em conjunto denotam que a rede possui uma boa conectividade entre os nodos.

Figura 3 - Rede de inter-relações de termos no contexto de *nanotechnology*



A seguir são detalhadas, para cada uma das comunidades, as principais informações que as compõem, sendo, os principais termos e suas respectivas medidas, as principais relações entre os termos em uma determinada comunidade e, por fim, o contexto geral indicando o foco da comunidade.

Na Comunidade 0 (cor roxa) consta um total de 67 termos que representam 46,53% de toda a rede. Levando-se em conta a centralidade de grau, os 5 (cinco) termos mais relevantes, *nanoparticle*, *monolayer*, *battery*, *graphene*, *electrical conductivity*, possuem valores de 123, 84, 81, 76, 52, respectivamente. Quanto à centralidade de proximidade (*closeness*) as ordens são mantidas se comparadas com a análise do grau, com exceção da quinta posição ocupada agora

pelo termo *electrical energy*, com valores de 0,877, 0,704, 0,694, 0,681, 0,611, respectivamente. Já para a centralidade de intermediação (*betweenness*) o termo mais relevante é *nanoparticle*, seguido pelos termos *graphene, battery, monolayer, energy storage* com valores de 1952,802, 885,472, 794,430, 648,828, 253,205, respectivamente. Analisando as relações percebe-se que os termos citados na avaliação das 3 (três) medidas possuem conexões com vários termos das demais comunidades. Aqui vale destacar dois outros termos, *energy density* e *energy storage*, que possuem, em sua maioria, conexões com termos da própria comunidade, com valores de centralidade de grau, proximidade e intermediação de 33, 0,556, 88,251, para *energy density*, e 40, 0,570, 253,205, para *energy storage*. Analisando-se a centralidade de intermediação do termo *energy storage* (253,205), este possui uma maior relevância quando comparado ao termo *electrical condutivity* (236,067). De modo geral, ao se verificar os termos pertencentes à comunidade pode-se afirmar que esta se refere à área de Energia, tanto pelos termos que trazem a designação "energy" e "electrical" na sua constituição, quanto termos em que as palavras "*power*" e "*light*" são mencionadas.

A Comunidade 1 (cor laranja) possui ao todo 60 termos representando 44,44% de toda a rede. Considerando a centralidade de grau, os cinco termos mais relevantes, *antibacterial, cancer cell, breast cancer, biocompatibility, drug delivery system*, possuem valores de 84, 84, 83, 81, 81, respectivamente. Com relação à centralidade de proximidade (*closeness*), referente a relevância dos termos, ocorre uma alteração entre os termos *drug delivery system e biocompatibility*, resultando em valores de 0,701, 0,701, 0,698, 0,694, 0,691, respectivamente. Já para a centralidade de intermediação (*betweenness*) o termo mais relevante é o *biocompatibility*, seguido pelos termos *breast cancer, lung cancer, chemotherapy, cancer cell*, com valores de 305,708, 247,244, 227,891, 211,811, 205,057, respectivamente. Analisando as relações percebe-se, assim como na Comunidade 0, que os termos citados possuem conexões com vários termos das demais comunidades. Além destes, destacam-se dois outros termos, *anticancer drug* e *combination therapy*, que possuem, em sua maioria, conexões com termos da própria comunidade, com valores de centralidade de grau, proximidade e intermediação de 43, 0,561, 4,556 para *anticancer drug*, e 49, 0,586, 5,926, para *combination therapy*. De modo geral, ao se verificar os termos pertencentes à comunidade pode-se afirmar que esta se refere à área Saúde, com foco no tratamento de câncer como pode ser constatado verificando-se os termos mais relevantes já mencionados.

Por fim, a Comunidade 2 possui um total de 13 termos que representam 9,03% de toda a rede. Considerando a centralidade de grau, os cinco termos mais relevantes, *methylene blue, reactive oxygen species, gram-positive bacteria, gram-negative bactéria* e *cytotoxic effects*, possuem valores de 58, 56, 44, 44, 34, respectivamente. Quanto à centralidade de proximidade (*closeness*) a relevância dos termos não se altera e estes mantêm a mesma ordem com valores de 0,619, 0,593, 0,563, 0,556, 0,536, respectivamente. Já para a centralidade de intermediação (*betweenness*) o termo mais relevante é o *methylene blue*, seguido pelos termos *reactive oxygen species, gram-positive bacteria, gram-negative bactéria e bacterial growth*, com valores de 65,623, 30,956, 10,408, 7,934, 6,418, respectivamente. Analisando as relações percebe-se que os termos citados em ambas as medidas possuem conexões com vários termos das demais comunidades. Todavia, os termos desta comunidade se conectam com maior frequência com termos que constam na Comunidade 1. Deste modo, analisando a comunidade pode-se afirmar que esta também se refere à área da Saúde com foco principalmente nos nodos que possuem em seus rótulos as palavras *antibacterial, antimicrobial e bacterial*.

Pensando no fluxo do método proposto torna-se relevante, após uma análise preliminar das interconexões dos termos que representam determinada temática, neste caso, *nanotechnology*, avaliar os documentos das patentes com o intuito de promover um entendimento mais específico. Desta forma, esta etapa visa promover uma ligação entre os termos mais relevantes encontrados em cada comunidade com as patentes. A seguir o Quadro 3 apresenta os quantitativos obtidos por meio de consultas à base de patentes. A variável *ti* refere-se ao termo de interesse *nanotechnology* (na consulta utiliza-se *nanotechnolog** para que sejam consideradas as derivações do termo).

Quadro 3 - Número de patentes referentes aos 5 (cinco) termos mais relevantes de cada comunidade considerando a medida de centralidade de grau

Comunidade	Conteúdo de Busca	Total de Patentes
Comunidade 0	*nanoparticle* (t_1) AND *ti*	434
	monolayer (t_2) AND *ti*	212
	battery (t_3) AND *ti*	282
	graphene (t_4) AND *ti*	269
	electrical conductivity (t_5) AND *ti*	166
	(t_1 OR t_2 OR t_3 OR t_4 OR t_5) AND *ti*	922
Comunidade 1	*antibacterial* (t_1) AND *ti*	135
	cancer cell (t_2) AND *ti*	169
	breast cancer (t_3) AND *ti*	163
	biocompatibility (t_4) AND *ti*	137
	drug delivery system (t_5) AND *ti*	80
	(t_1 OR t_2 OR t_3 OR t_4 OR t_5) AND *ti*	400
Comunidade 2	*methylene blue* (t_1) AND *ti*	51
	reactive oxygen species (t_2) AND *ti*	51
	gram-positive bacteria (t_3) AND *ti*	31
	gram-negative bacteria (t_4) AND *ti*	40
	cytotoxic effects (t_5) AND *ti*	23
	(t_1 OR t_2 OR t_3 OR t_4 OR t_5) AND *ti*	128

A análise do texto das patentes recuperados a partir dos termos mais relevantes de cada comunidade permite ainda avançar no entendimento do contexto de interesse. Por exemplo, ao se analisar as patentes torna-se possível identificar novos termos com o intuito de aprimorar o dicionário de termos. A partir disso, ao se gerar novamente determinada rede, conexões adicionais serão estabelecidas permitindo uma visão mais ampla dos conceitos, técnicas e tecnologias que compõem o contexto de interesse. No Quadro 4 são apresentados os títulos das 10 patentes mais relevantes através de consulta com os 5 termos mais relevantes da Comunidade 0. A consulta é efetuada no título, no resumo e no texto completo da patente. Por questões de simplificação e demonstração somente o título da patente é apresentado.

Quadro 4 - Relação de títulos de patentes.

Id	Título da Patente
1	*Graphene powder, method for producing graphene powder and electrode for lithium ion battery containing graphene powder*
2	*Electrode material comprising graphene-composite materials in a graphite network*
3	*Graphene dispersion pastes, methods of preparing and using the same*
4	*Integration of monolayer graphene with a semiconductor device*
5	*Real-time detection of water contaminants*
6	*Micronized composite powder additive*
7	*Magnetic graphene like nanoparticles or graphitic nano- or microparticles and method of production and uses thereof.*
8	*Process for forming a nanocomposite film.*
9	*Method for making polyvinyl alcohol/carbon nanotube nanocomposite film.*
10	*Nanocomposite film comprising cellulose and a conductive nanofiller, and method of making.*

No quadro acima vários termos encontram-se destacados. A cor amarela indica que o termo já consta na lista original utilizada neste estudo, enquanto que a cor verde também indica um termo que consta na lista original, mas escrito de maneira diferente, ou seja, *li-ion battery*. Nesta situação o termo será trato como um sinônimo para a busca. Já a cor azul os termos novos que podem ser adicionados ao dicionário. A mesma análise pode ser efetuada para as demais comunidades possibilitando o aprimoramento do dicionário de termos. No método atual, o conteúdo das patentes é apresentado ao especialista que tem a função de identificar os termos que serão adicionados ao dicionário como novos ou sinônimos.

O método fornece passos que podem auxiliar na investigação e entendimento de redes de tecnologia que se formam a partir de um termo de interesse. Sendo assim, promove subsídios para que especialistas, em seus departamentos de Pesquisa, Desenvolvimento e Inovação (PD&I), possam avaliar adequadamente determinada área buscando determinar caminhos para maximizar e definir investimentos. Ademais, o método surge como uma ferramenta com foco na gestão da inovação, uma vez que, a análise de patentes tem potencial para melhorar ou definir novos produtos e processos impactando na competitividade de organizações de base tecnológica.

4. CONCLUSÃO

Com a tarefa de expor a relevância da análise de inter-relação de termos, em específico, a partir de uma base de patentes, o estudo realizado com o intuito de avaliar o método proposto se apresenta adequado. Isto decorre da identificação de termos fortemente relacionados, termos de maior influência, bem como a formação de grupos que habilitam verificar possíveis lacunas em um domínio de conhecimento. A implementação do método foi exemplificada com uma aplicação na área da nanotecnologia. Para tal, utilizou-se a base de patentes americanas referente aos anos de 2019 e 2020, sendo que a rede de termos produzida evidencia o interesse de aplicações de nanotecnologia para as áreas da saúde (Comunidades 1 e 2) e energia (Comunidade 0).

No tocante às limitações do método, a elaboração do dicionário de termos para versões futuras poderá ser automatizado por extrações via NER (*Name-Entity Recognition*). Assim como, se tornar iterativo adicionando novos termos ao dicionário a partir da exploração dos documentos de patentes que o modelo agrupa. Neste sentido, a análise se torna cíclica e o conhecimento sobre o domínio de interesse, representado pela temática, sofre constantes evoluções. No que tange a visualização de informação, outras formas de representação, tais como, gráficos, *tag clouds*, árvores, podem ser utilizadas permitindo um aprofundamento na temática de interesse.

Referente às redes apresentadas neste estudo, estas foram elaboradas a partir de uma análise que considera a frequência, número de documentos de patentes nos quais havia coocorrência dos termos. No entanto, a rede pode ser explorada considerando a correlação direta ou a associação (relação indireta) de elementos textuais (termos). Além disso, pretende-se, no avanço do método, implementar conceitos de temporalidade na análise para investigar a evolução das relações. Por fim, levando-se em conta a temática de interesse utilizada, a nanotecnologia, se estende a muitos campos da ciência, engenharia e tecnologia, o que torna desafiador seu mapeamento e organização.

AGRADECIMENTO

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001. Agradecemos a Universidade do Estado de Mato Grosso pelo apoio através de qualificação docente.

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

Barabási, A.-L. (2013). Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1987), 20120375. https://doi.org/10.1098/rsta.2012.0375

Barabási, A.-L., & Albert, R. (1999). Emergence of Scaling in Random Networks. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509

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

Chang, S.-H. (2018). A pilot study on the connection between scientific fields and patent classification systems. *Scientometrics*, *114*(3), 951–970. https://doi.org/10.1007/s11192-017-2613-6

Erdös, P.; Rényi, A. (1959). On random graphs, i. *Publicationes Mathematicae (Debrecen)*, *6*, 290–297.

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

Fanfair, D., Desai, S., & Kelty, C. (2007). The early history of nanotechnology. *Connexions*, *6*. Retrieved from https://cnx.org/contents/Altp_xOu@1/The-Early-History-of-Nanotechnology

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*, *40*(1), 35. https://doi.org/10.2307/3033543

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

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, *99*(12), 7821–7826. https://doi.org/10.1073/pnas.122653799

Grames, E. M., Stillman, A. N., Tingley, M. W., & Elphick, C. S. (2019). An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods in Ecology and Evolution*, *10*(10), 1645–1654. https://doi.org/10.1111/2041-210X.13268

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

Jiang, S., Gao, Q., Chen, H., & Roco, M. C. (2015). The roles of sharing, transfer, and public funding in nanotechnology knowledge-diffusion networks. *Journal of the Association for Information Science and Technology*, *66*(5), 1017–1029. https://doi.org/10.1002/asi.23223

Jun, S., & Park, S. (2016). Examining technological competition between BMW and Hyundai in the Korean car market. *Technology Analysis & Strategic Management*, *28*(2), 156–175. https://doi.org/10.1080/09537325.2015.1073252

Kagan, C. R., Fernandez, L. E., Gogotsi, Y., Hammond, P. T., Hersam, M. C., Nel, A. E., … Weiss, P. S. (2016). Nano Day: Celebrating the Next Decade of Nanoscience and Nanotechnology. *ACS Nano*, *10*(10), 9093–9103. https://doi.org/10.1021/acsnano.6b06655

Katsurai, M., & Ono, S. (2019). TrendNets: mapping emerging research trends from dynamic co-word networks via sparse representation. *Scientometrics*, *121*(3), 1583–1598. https://doi.org/10.1007/s11192-019-03241-6

Li, Q., Zhang, H., & Hong, X. (2020). Knowledge structure of technology licensing based on co-keywords network: A review and future directions. *International Review of Economics & Finance*, *66*, 154–165. https://doi.org/10.1016/j.iref.2019.11.007

Liu, J., Tang, T., Wang, W., Xu, B., Kong, X., & Xia, F. (2018). A Survey of Scholarly Data Visualization. *IEEE Access*, *6*, 19205–19221. https://doi.org/10.1109/ACCESS.2018.2815030

Madani, F., & Weber, C. (2016). The evolution of patent mining: Applying bibliometrics analysis and keyword network analysis. *World Patent Information*, *46*, 32–48. https://doi.org/10.1016/j.wpi.2016.05.008

Molontay, R., & Nagy, M. (2019). Two decades of network science. *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 578–583. https://doi.org/10.1145/3341161.3343685

Muñoz-Écija, T., Vargas-Quesada, B., & Chinchilla-Rodríguez, Z. (2017). Identification and visualization of the intellectual structure and the main research lines in nanoscience and

nanotechnology at the worldwide level. *Journal of Nanoparticle Research*, *19*(2), 62. https://doi.org/10.1007/s11051-016-3732-3

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, Vol. 45, pp. 167–256. https://doi.org/10.1137/S003614450342480

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

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*. Retrieved from http://www.nano.gov/sites/default/files/nki_nsi_white_paper_-_may_14_2012_secured.pdf

Ozcan, S., & Islam, N. (2014). Collaborative networks and technology clusters - The case of nanowire. *Technological Forecasting and Social Change*, *82*(1), 115–131. https://doi.org/10.1016/j.techfore.2013.08.008

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

Radhakrishnan, S., Erbis, S., Isaacs, J. A., & Kamarthi, S. (2017). Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature. *PLOS ONE*, *12*(3), e0172778. https://doi.org/10.1371/journal.pone.0172778

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

Su, H. N., & Lee, P. C. (2010). Mapping knowledge structure by keyword co-occurrence: A first look at journal papers in Technology Foresight. *Scientometrics*, *85*(1), 65–79. https://doi.org/10.1007/s11192-010-0259-8

Toumey, C. (2009). Plenty of room, plenty of history. *Nature Nanotechnology*, Vol. 4, pp. 783–784. https://doi.org/10.1038/nnano.2009.357

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing & Management*, *43*(5), 1216–1247. https://doi.org/https://doi.org/10.1016/j.ipm.2006.11.011

Vespignani, A. (2018). Twenty years of network science. *Nature*, *558*(7711), 528–529. https://doi.org/10.1038/d41586-018-05444-y

Wang, G., & Guan, J. (2011). Measuring science-technology interactions using patent citations and author-inventor links: An exploration analysis from Chinese nanotechnology. *Journal of Nanoparticle Research*, *13*(12), 6245–6262. https://doi.org/10.1007/s11051-011-0549-y

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, *393*(6684), 440–442. https://doi.org/10.1038/30918

Yang, B., Liu, D., & Liu, J. (2010). Discovering Communities from Social Networks: Methodologies and Applications. In *Handbook of Social Network Technologies and Applications* (pp. 331–346). https://doi.org/10.1007/978-1-4419-7142-5_16

Zingg, R., & Fischer, M. (2018). The nanotechnology patent thicket revisited. *Journal of Nanoparticle Research*, *20*(10). https://doi.org/10.1007/s11051-018-4372-6

Zingg, R., & Fischer, M. (2019). The consolidation of nanomedicine. *Wiley Interdisciplinary*

Reviews: Nanomedicine and Nanobiotechnology, *11*(6), 1–6. https://doi.org/10.1002/wnan.1569

APÊNDICE I

Lista dos 257 termos e palavras-chave da área de nanotecnologia utilizadas no estudo

A549 cell	cancer chemotherapy	differential pulse voltammetry	freundlich
Ag NPs	cancer diagnosis	discharge capacity	functional monomer
aldehydes	cancer therapy	dna damage	galvanostatic charge/discharge
alkaline media	cancer treatment	dose-dependent manner	glutathione
anode material	candida albicans	doxorubicin	gold nanoparticle
anode material	capacity fading	drug delivery system	gram-negative bacteria
antibacterial	capacity retention	drug resistance	gram-positive bacteria
antibacterial activity	carbon coating	dye adsorption	graphene
anticancer activity	carbon dots	dye degradation	graphene quantum dots
anticancer drug	caspase-3	dye-sensitized solar cell	graphitic carbon nitride
anti-inflammatory	catalytic reduction	early diagnosis	green solvent
antimicrobial	cathode material	eco-friendly	H-2 evolution
antimicrobial activity	cell apoptosis	efficient adsorbent	H-2 production
antimicrobial property	cell viability	electric vehicle	half-wave potential
antioxidant	cellular internalization	electrical conductivity	harsh condition
aptamer	cellular uptake	electrical energy	heat transfer characteristics
aquatic environment	cellulose nanocrystals	electrocatalytic performance	heat transfer coefficient
aromatic aldehydes	charge recombination	electrochemical energy storage	heat transfer performance
aryl halides	charge transfer resistance	electrochemical performance	heat transfer rate
asymmetric supercapacitor	chemotherapy	electrochemical sensor	heavy metal
bacterial growth	combination therapy	electrochemical signal	heavy metal ions
base fluid	conductive additive	electrochemical stability	hela cell
batch adsorption experiment	core shell structure	electrode material	hepatocellular carcinoma
batch experiment	coulombic efficiency	electrolyte ions	HepG2 cell
battery	counter electrode	energy conversion	heterogeneous catalyst
BET analysis	Cr(VI)	energy density	hexagonal boron nitride
biocompatibility	crude oil	energy efficiency	hierarchical porous structure
biofilm formation	curcumin	energy storage	high adsorption capacity
bioimaging	current collector	energy-dispersive x-ray	high biocompatibility
biological fluids	cycle stability	environmental remediation	high capacitance
biomass	cycling performance	equilibrium data	high energy density
biomedicine	cycling stability	excellent conductivity	high power density
biosynthesis	cytocompatibility	excellent photocatalytic activity	high rate capability
black phosphorus	cytotoxic effects	excellent selectivity	high theoretical capacity
box-behnken design	cytotoxicity	fesem image	high toxicity
breast cancer	degradation efficiency	folic acid	higher cytotoxicity
cancer cell	different weight ratio	free dox	higher specific capacitance

high-performance liquid chromatography	low electrical conductivity	nanofluid	photogenerated charge
high-performance supercapacitor	low limit	nanomedicine	photogenerated electron
human health	low overpotential	nanoparticle	photothermal therapy
hyaluronic acid	low toxicity	nanorobot	photovoltaic
hydrogen evolution reaction	lung cancer	nanosheet	physicochemical
hydrogen generation	magnetic nanoparticle	nir irradiation	plasmonic nanoparticle
hydroxyl radical	magnetic separation	nir light	plasmonic nanostructure
improved photocatalytic activity	malachite green	nusselt number	p-nitrophenol
langmuir adsorption isotherm	maximum adsorption capacity	open circuit voltage	poly (3,4-ethylenedioxythiophene)
langmuir isotherm	maximum power density	oral bioavailability	polydopamine
langmuir isotherm model	mcf-7	oral delivery	polysulfides
langmuir model	melamine	ordinary differential equations	polyvinylpyrrolidone
large specific surface area	mesenchymal stem cells	organic pollutants	poor cycling stability
large surface area	metal nps	orr activity	porous medium
large volume change	metal-air battery	osteogenic differentiation	power conversion efficiency
large-scale application	metal-organic frameworks	oxidative stress	power density
leading cause	metastasis	oxygen evolution reaction	prepared catalyst
leaf extract	methyl orange	oxygen reduction reaction	prepared composite
light absorption	methylene blue	oxygen-containing functional group	pristine graphene
lignin	mild reaction condition	paclitaxel	protein corona
li-ion battery	modulation depth	pathogenic bacteria	pseudomonas aeruginosa
linear dynamic range	molybdenum disulfide	photoanode	pseudo-second-order kinetic
lithiation	molybdenum disulfide	photocatalysis	reactive oxygen species
lithium-sulfur battery	monolayer	photocatalytic activity	styrenesulfonate
long cycle life	monolayer graphene	photocatalytic reduction	surface plasmon resonanc
low bioavailability	na-ion battery	photocatalytic water	
low cytotoxicity	nanocarrier	photocurrent density	
low detection limit	nanocatalyst	photoelectrochemical	