Uma proposta de metodologia de pesquisa de patentes verdes utilizando padrões frequentes

[1]CEFET/RJ - Centro Federal de Educação Tecnológica Celso Suckow da Fonseca
Av. Maracanã, 229 - Rio de Janeiro - RJ - Brasil.

Abstract. *With technological innovations advancement, a large number of documents are being generated in the form of patent registrations. Thus, the need for new techniques and tools capable of organizing and analyzing this data to present them with satisfactory performance increases. This paper aims to present a new approach to patent research and to make comparisons with methods studied today. Specifically, it conducts a study using a subset of patents called "green patents", developing a methodology for researching green patents using frequent pattern mining. We also present the expected contributions to the end of this work.*

Resumo. *Com o avanço das inovações tecnológicas, vem sendo gerada uma grande quantidade de documentos na forma de registros de patentes. Desta forma, cresce a necessidade de se obter novas técnicas e ferramentas capazes de organizar e analizar esses dados para apresentá-los com uma performance satisfatória. Este trabalho tem por objetivo apresentar uma nova abordagem de pesquisa de patentes e realizar comparações com métodos estudados nos dias de hoje. Especificamente, realiza um estudo utilizando um subconjunto de patentes denominado "patentes verdes", desenvolvendo uma metodologia para pesquisa de patentes verdes utilizando mineração de padrões frequentes. Também são apresentadas as contribuições esperadas ao fim deste trabalho.*

1. Introdução

O registro de patentes serve como um indicador de atividades econômicas e tecnológicas, motivando o desenvolvimento tecnológico e de ideias. Isso se deve pela exclusividade de lucro que um inventor passa a ter como direito pelo registro de suas invenções [Ernst, 2003; Almeida and Kogut, 1997]. Questões de competitividade, inovação e custos contribuem como estímulo às empresas em desenvolverem tecnologia e registrar novas patentes, o que vem resultando num grande aumento da quantidade de registro de patentes, necessitando sistemas eficazes e eficientes para administrar essa grande quantidade de dados [Shalaby and Zadrozny, 2017].

Além dessas questões, outros fatores acabam por influenciar um aumento desse desenvolvimento de tecnologias em áreas específicas. Tratados internacionais sobre meio ambiente vêm causando pressões regulatórias e normativas sobre questões ambientais, o que acaba por influenciar positivamente empresas a buscarem soluções que causem menor impacto ambiental [Berrone et al., 2013]. Um notório exemplo disso é a política de transição energética da Alemanha (*Energiewende*), que foi responsável por mais que

triplicar o número absoluto de patentes alemãs de energia eólica entre 2005 e 2010, aumentando em um quarto o número de patentes de energia solar fotovoltaica no mesmo período [Pegels and Lütkenhorst, 2014].

Com os volumes cada vez maiores de informações de patentes, as tarefas de busca e análise de patentes tornaram-se vitais sob perspectivas legais e gerenciais [Liu et al., 2011]. Isso mostra importância de metodologias e ferramentas desenvolvidas para analizar dados de patentes e apresentá-los com uma performance satisfatória, tais como *Patent retrieval*.

Patent retrieval é considerado uma ferramenta fundamental para a maioria das tarefas de analise de patentes. Seu desenvolvimento foi motivado por alguns problemas específicos ao tentar aplicar técnicas tradicionais de mineração de texto à análise de patentes [Shalaby and Zadrozny, 2017]. Dentre eles, está o fato de que registros de patentes são combinações de texto descritivo **não-estruturados** (e.g., "Título da Patente", "Resumo da Patente" e a "Descrição da Patente"), e dados bibliográficos **estruturados** relacionados a cada patente (e.g., "Inventores", "Concessionários" e "Citações") [Liu et al., 2011] .

Apesar da importância do tema, ainda é uma área que carece de estudos desenvolvendo metodologias a serem empregadas na tarefa, sendo encontrados poucos trabalhos que tratam de *Patent retrieval* [Abbas et al., 2014; Shalaby and Zadrozny, 2017]. Além disso, os resultados em tarefas de *Patent retrieval* apresentam desempenho mediano, mesmo nos trabalhos apontados como estado da arte [Shalaby and Zadrozny, 2017]. Considerando este cenário, essa proposta de pesquisa tem por objetivo obter melhores resultados em tarefas de *Patent retrieval*, desenvolvendo para isso uma combinação de abordagens que usam conceitos de mineração de textos, associadas às técnicas de mineração de padrões frequentes.

O restante desse trabalho está organizado da seguinte forma: na Seção 2, são apresentados conceitos sobre temas abordados neste estudo, de forma a falar sobre mineração de texto, especificamente *Patent retrieval*, detalhando também os conceitos de mineração de padrões frequentes. Na Seção 3, é descrita a metodologia proposta neste estudo. Na Seção 4, são relacionados trabalhos referentes à analise e apresentação de dados de patentes. Por fim, na Seção 5, há uma discussão sobre as conclusões e cenários futuros.

2. Conceitos

2.1. Mineração de Patentes

Patentes são a concessão de um título de propriedade temporária de invenções de produtos, equipamentos, ferramentas e até mesmo de procedimentos e métodos de negócio [Mansfield, 1986].

Madani and Weber [2016], existem três estágios principais da evolução da mineração de patentes. No primeiro estágio, as metodologias mais utilizadas eram a análise bibliométrica e a análise de citações. O segundo estágio foi caracterizado pelo maior uso da análise de *clusters* e de redes. No terceiro estágio a mineração de patentes evoluiu para métodos de análise que utilizam a mineração de textos e outras metodologias complementares, como análise semântica e análise ontológica.

Liu et al. [2011] apresenta um fluxo genérico de analise de patente detalhado como pode ser observado na Figura 1.

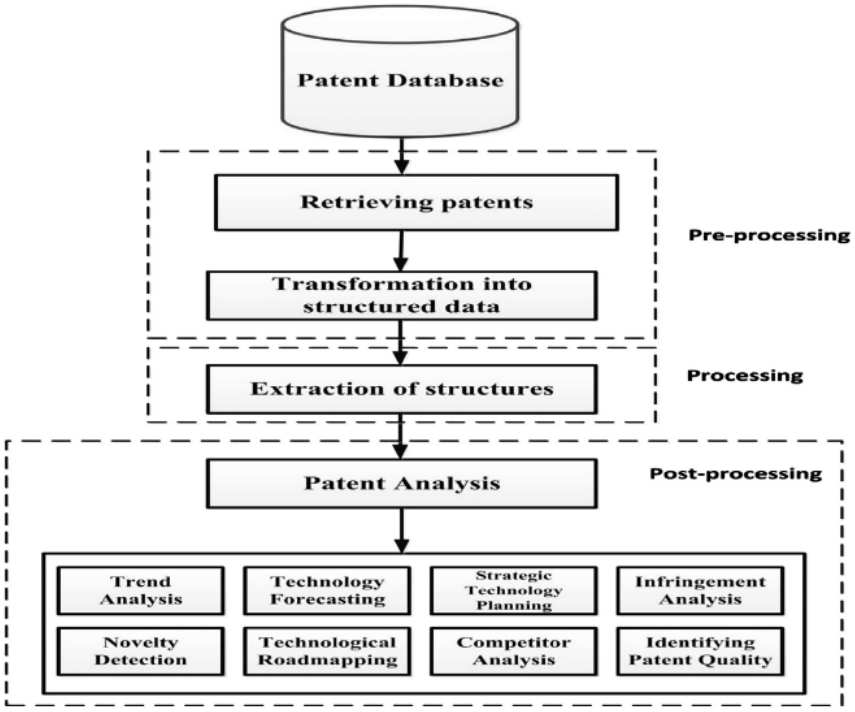

Figura 1. Fluxo genérico de Análise de Patente (adaptado de Abbas et al. [2014])

2.2. Mineração de Padrões Frequentes

O processo conhecido como mineração de dados consiste em explorar grandes quantidades de dados, de forma a extrair ou ajudar na obtenção de informações interessantes [Han et al., 2011]. O uso de técnicas da área de mineração de dados vem atender a necessidade de análise destes dados, considerando a grande quantidade de dados sendo gerados e armazenados a todo momento na atualidade [Chen et al., 2014]. Essa área é focada na automatização da análise de grandes conjuntos de dados, gerando novos subconjuntos organizados de forma a encontrar padrões consistentes, destacando-se técnicas conhecidas como regras de associação, análise de sequências temporais e mineração de padrões frequentes [Han et al., 2011].

A mineração de padrões frequentes é um método para extrair conhecimento de dados buscando padrões recorrentes. Dentre diversas técnicas para mineração de padrões frequentes, pode ser utilizado o algoritmo Apriori [Agrawal et al., 1994]. Esse algoritmo busca por padrões e encontra conjuntos de itens frequentes, podendo resultar em conhecimento novo e útil sobre associações e correlações entre dados [James et al., 2013].

Ao concluir a execução de técnica para mineração de padrões frequentes, apresenta-se como resultado algo na forma: “SE X, ENTÃO Y”. Ou seja, em uma mesma transação aparece uma representação de um ou mais elementos que levam à presença de outro elemento ou elementos (‘transação’ indicando, aqui, quais itens foram consultados em uma determinada operação de consulta). Com isso se encontram relacionamentos ou padrões frequentes, como conjuntos de itens frequentes, subsequencias frequentes e subestruturas frequentes, entre conjuntos de dados [Han et al., 2011].

Padrões que refletem itens que são frequentemente associados podem ser repre-

sentados sob a forma de regras de associação. Essas regras possuem medidas chamadas de suporte e confiança. O suporte representa a proporção de transações em análise em que os itens identificados aparecem juntos. Confiança representa a proporção que dado item aparece junto de outro, dentro do subconjunto em que dado item sempre aparece. As regras de associação que são consideradas interessantes são aquelas que alcançam os valores mínimos de suporte e confiança.

Porém, mesmo com a delimitação de valores mínimos para suporte e confiança, pode acontecer de ser encontrado um número enorme de regras geradas. Isso acontece principalmente quando os valores mínimos de suporte e confiança são muito baixos. Assim, é utilizada também uma medida chamada *lift* (elevação) [Webb, 2000], como forma de análise adicional para selecionar regras mais relevantes.

3. Metodologia Proposta

Conforme ilustrado na Figura 2, a metodologia para desenvolvimento deste trabalho se baseia, inicialmente, em analisar um conjunto de dados ("conjunto de treinamento") e organizá-los de acordo com padrões. O reconhecimento de padrões visa classificar dados baseados em conhecimento *a priori* (i.e., preliminar ou dedutivo) ou informações estatísticas extraídas de padrões. Em seguida, usar matriz de similaridade e realizar uma técnica de mineração de textos (*textmining*) para comparação, e então avaliar o melhor método que traz os resultados. Replicar os métodos que são mais utilizados para gerar dados para comparação dos métodos.

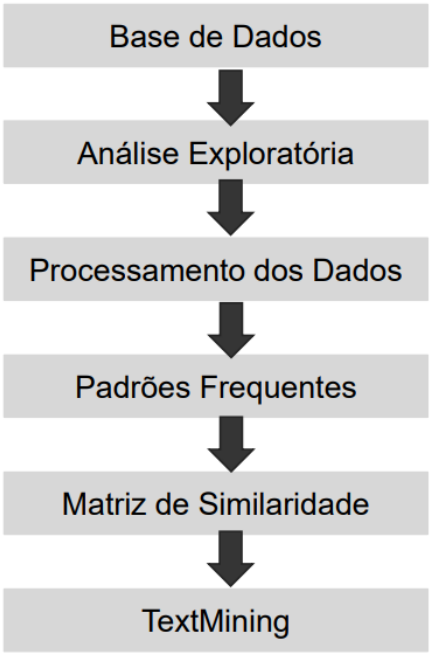

Figura 2. Sequência de tratamento dos dados

Serão aplicadas técnicas de pré-processamento, antes de se buscar por padrões frequentes nos dados de patentes obtidos. Considerando as etapas encontradas na literatura, são previstas as tarefas de integração, redução e transformação dos dados utilizados. O

objetivo desta etapa é de aumentar a qualidade e o poder de expressão dos dados a serem minerados [Han et al., 2011]. Em seguida, será dada sequência à busca por padrões frequentes.

Referente à metodologia utilizada para busca de padrões frequentes em dados de patentes, o processo será baseado nas atividades de geração de regras e análise de regras. Na geração de regras é aplicado um princípio *a priori* para produzir regras de associação. Podem ser atribuídos valores de suporte e a confiança de $0,01$ e $0,5$, respectivamente, como forma de limitar o número de regras obtidas. Estudos de trabalho na literatura serão buscados como subsídio para cálculo do valor utilizado, de forma a reduzir a quantidade de regras a serem interpretadas durante a análise de regras sem prejuízo para a tarefa.

A aprendizagem de regras de associação envolve técnicas de análise destinadas a descobrir as associações e conexões entre itens específicos. Para geração de regras de regras de associação, se utiliza o algoritmo Apriori, por exemplo. Proposto em 1994, trata-se de um algoritmo que resolve o problema da mineração de *itemsets* frequentes. Ou seja, dados um banco de dados de transações D e um nível mínimo de suporte S, o algoritmo encontra todos os *itemsets* frequentes com relação a D e S [Agrawal et al., 1994].

Após a análise de padrões será obtido um agrupamento com a intenção de diminuir o escopo da massa de dados onde espera-se obter um melhor desempenho na realização de busca. Após esta o agrupamento serão preenchidos os dados numa matriz de similaridade.

4. Trabalhos relacionados

Liu et al. [2011] desenvolveram um sistema de recuperação de patentes que integra o acoplamento bibliográfico com a mineração de texto para melhorar a recuperação de patentes. Na solução que propuseram, obtiveram uma redução do número de patente recomendada para mais de 90% usando a melhor combinação. Apontam no trabalho que foi identificada a necessidade de mais testes, envolvendo especialistas da área de patentes para ajudar a melhorar a eficácia.

Com o foco em aumentar transferência tecnológica, Park et al. [2013b] propuseram uma metodologia para identificação de tecnologias de alto valor. Para isso, adotaram as tendências TRIZ, classificadas considerando as características da etapa do ciclo de vida, como critérios para avaliar tecnologias em patentes. Eles verificaram o método aplicando-o à tecnologia de turbinas eólicas flutuantes, uma tecnologia relacionada à área de patentes verdes.

Yoon et al. [2013] construíram mapas dinâmicos de patentes para mostrar as tendências da concorrência tecnológica e descrevem as funções estratégicas dos mapas dinâmicos. Na trabalho, eles introduziram a análise semântica de patentes com vantagens na representação de objetivos e estruturas tecnológicas.

A proposta de [Park et al., 2013a] cria um sistema denominado TechPerceptor focado na análise de patente de forma mais inteligente. O mesmo realiza uma análise gramatical usando conceito processamento de Linguagem Natural da área de *textmining* individualmente para cada patente, através de uma analise semântica da estrutura são construídos uma rede e um mapeamento das patentes buscando similaridade entre as mesmas.

A pesquisa realizada por Trappey et al. [2013] utiliza o algoritimo de Back propagation e foca em minimizar os esforços e o tempo necessário para procurar e determinar a qualidade da patente para gerenciar as operações de P&D especialmente para uma inovação.O propósito de treinar através de um algoritimo de back-propagation é identificar as patentes que são específico para uma tecnologia e fazer uma recomendação precisa. As patentes identificadas são então classificadas para ajudar a entender valor técnico das patentes.

Choi et al. [2012] propõem uma abordagem baseada em mineração de texto que desenvolve uma Árvore Tecnológica (TechTree), explorando e examinando informações sobre patentes. A informação extraída através das estruturas é categorizada com base em semelhanças. Dois importantes processos da abordagem proposta são: (i) o desenvolvimento procedimentos para construir dados de origem de patentes; e (ii) um método para construir uma TechTree a partir desses dados. O processamento de linguagem natural é usado para extrair as estruturas e técnicas de mineração de texto são usadas para análise das mesmas. As semelhanças entre as estruturas das patentes são calculadas e uma matriz de similaridade é produzida.

Gerken and Moehrle [2012] adotaram uma abordagem de análise semântica para identificar invenções em patentes que são altamente novas. O primeiro passo da abordagem proposta extrai estruturas semânticas a partir dos dados de patentes textuais. As estruturas semânticas são extraídas através da análise sintática das patentes, utilizando marcação de trechos dos dados. Em seguida, estruturas semânticas são identificadas e análise linguística particular para elementos relacionados a domínio e situação são executados. A análise é importante para resolver os problemas de sinônimos que podem surgir do domínio ou situação específica. Na terceira etapa, a medida de similaridade é realizada e uma matriz de similaridade é criada com base na comparação de estruturas semânticas. Com a matriz de similaridade construída, a novidade da patente é determinada pela comparação do valores da matriz.

Na Tabela 1 é possível observar a relação entre as propostas abordadas nos trabalhos relacionados e as técnicas apresentadas pelos mesmos.

5. Considerações

A pesquisa apresentada neste trabalho se encontra em uma fase inicial de desenvolvimento. Ainda será implementada a etapa onde serão obtidos resultados preliminares com a implementação de padrões frequentes. A expectativa é que sejam obtidos resultados melhores do que utilizando técnicas descritas nos trabalhos relacionados.

Considerando que as bases de dados disponíveis são extensas, a utilização de técnicas como matriz de similaridade, sem o uso de ferramentas que possam ajustar esses dados previamente [Gerken and Moehrle, 2012], pode exigir bastante durante processamento dos dados e na obtenção de boas métricas de resultados. Mesmo assim, os resultados de trabalhos relacionados indicam que técnicas de mineração de patentes podem apresentar melhorias, caso aja a combinação de técnicas de recuperação de informação a partir de dados estruturados e não-estruturados. Isso é um fator que, associado à importância do tema, abre caminho para futuros trabalhos nessa linha de estudo e estimula que sejam realizados novos experimentos.

Um grande desafio que se apresenta é o de trabalhar com a grande quantidade de

Tabela 1. Comparação entre técnicas de analise de Patentes

Trabalho[Ano]	Proposta	Técnica aplicada
Liu et al. [2011]	Melhoria da precisão de pesquisa, identificação de similaridade	*textmining* , acoplamento bibliográfico
Park et al. [2013b]	Identificação de patentes promissoras	*textmining*
Yoon et al. [2013]	Identificação de vácuos tecnológicos, hotspots tecnológicos e tendências	Mapeamento de patentes, Processamento de linguagem natural
Park et al. [2013a]	Extração de informação particular para cada funcionalidade da patente	Processamento de linguagem natural
Choi et al. [2012]	Desenvolvimento de ferramenta de busca em Árvore (TechTree)	Processamento de linguagem natural, *textmining*
Trappey et al. [2013]	Determinação de qualidade de patentes para operações de P&D	Algorítimo de Back-propagation
Gerken and Moehrle [2012]	Detecção de novidade	Detecção de Similaridade

dados na realização da comparação da matriz de similaridade. Nesse caso, dependendo do tamanho da base utilizada, pode ser gerada uma matriz de similaridade muito grande no processo de comparação usando o *textmining*. Desta forma, são esperadas situações onde o tempo para processar esses dados seja muito grande, o que justifica a alocação de uma máquina com um poder de processamento suficientemente alto para tratar essas informações em um tempo que seja considerado viável.

Uma das ideias a serem experimentadas para contornar as dificuldades esperadas é de realizar o cálculo de forma individual e armazenar no banco de dados para consulta dos resultados a posteriori. Ainda não foram esgotadas as pesquisas para subsidiar as decisões que dizem respeito a rodar padrões frequentes para verificar se o agrupamento produz uma vantagem satisfatória no tamanho dos dados.

Referências

Abbas, A., Zhang, L., and Khan, S. U. (2014). A literature review on the state-of-the-art in patent analysis. *World Patent Information*, 37:3–13.

Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

Almeida, P. and Kogut, B. (1997). The exploration of technological diversity and geographic localization in innovation: Start-up firms in the semiconductor industry. *Small Business Economics*, 9(1):21–31.

Berrone, P., Fosfuri, A., Gelabert, L., and Gomez-Mejia, L. R. (2013). Necessity as the mother of 'green'inventions: Institutional pressures and environmental innovations. *Strategic Management Journal*, 34(8):891–909.

Chen, M., Mao, S., and Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, 19(2):171–209.

Choi, S., Park, H., Kang, D., Lee, J. Y., and Kim, K. (2012). An sao-based text mining approach to building a technology tree for technology planning. *Expert Systems with Applications*, 39(13):11443–11455.

Ernst, H. (2003). Patent information for strategic technology management. *World patent information*, 25(3):233–242.

Gerken, J. M. and Moehrle, M. G. (2012). A new instrument for technology monitoring: novelty in patents measured by semantic patent analysis. *Scientometrics*, 91(3):645–670.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Liu, S.-H., Liao, H.-L., Pi, S.-M., and Hu, J.-W. (2011). Development of a patent retrieval and analysis platform–a hybrid approach. *Expert systems with applications*, 38(6):7864–7868.

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

Mansfield, E. (1986). Patents and innovation: an empirical study. *Management science*, 32(2):173–181.

Park, H., Kim, K., Choi, S., and Yoon, J. (2013a). A patent intelligence system for strategic technology planning. *Expert Systems with Applications*, 40(7):2373–2390.

Park, H., Ree, J. J., and Kim, K. (2013b). Identification of promising patents for technology transfers using triz evolution trends. *Expert Systems with Applications*, 40(2):736–743.

Pegels, A. and Lütkenhorst, W. (2014). Is germany s energy transition a case of successful green industrial policy? contrasting wind and solar pv. *Energy Policy*, 74:522–534.

Shalaby, W. and Zadrozny, W. (2017). Patent retrieval: a literature review. *Knowledge and Information Systems*, pages 1–30.

Trappey, A. J., Trappey, C. V., Wu, C.-Y., Fan, C. Y., and Lin, Y.-L. (2013). Intelligent patent recommendation system for innovative design collaboration. *Journal of Network and Computer Applications*, 36(6):1441–1450.

Webb, G. I. (2000). Efficient search for association rules. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–107. ACM.

Yoon, J., Park, H., and Kim, K. (2013). Identifying technological competition trends for r&d planning using dynamic patent maps: Sao-based content analysis. *Scientometrics*, 94(1):313–331.