Exploração de dados e visualização de informação

Documento descritivo de soluções em software para análises de dados de serviços de informações sobre patentes

Centro de Gestão e Estudos Estratégicos
Ciência, Tecnologia e Inovação

Exploração de dados e visualização de informação

Documento descritivo de soluções em software para análises de dados de serviços de informações sobre patentes

Brasília-DF,
Dezembro, 2022

Centro de Gestão e Estudos Estratégicos (CGEE)

Organização social supervisionada pelo Ministério da Ciência, Tecnologia, Inovações (MCTI)

Documento descritivo de soluções em software para análises de dados de serviços de informações sobre patentes. Exploração de dados e visualização de informação. Brasília, DF: Centro de Gestão e Estudos Estratégicos, 2022.

13p.

1. Ciência de dados. 2. Patentes. 3. Inovação.
I. Título. II. CGEE. III. MCTI.

Centro de Gestão e Estudos Estratégicos - CGEE
SCS Quadra 9 – Torre C – 4º andar – salas 401 a 405
Edifício Parque Cidade Corporate
70308-200 - Brasília, DF
Telefone: (61) 3424.9600
http://www.cgee.org.br

Referência bibliográfica:
Centro de Gestão e Estudos Estratégicos - CGEE. **Documento descritivo de soluções em software para análises de dados de serviços de informações sobre patentes.** Exploração de dados e visualização de informação.. Brasília, DF: 2022. 23p.

Este documento é parte integrante das atividades desenvolvidas no âmbito do 2º Contrato de Gestão CGEE – 38º Termo Aditivo/Projeto: Apoio técnico para desenvolvimento de ações de avaliação no CNPq – 8.10.56.01.51.01/ Ministério da Ciência, Tecnologia e Inovações/2022.

Exploração de dados e visualização de informação

Documento descritivo de soluções em software para análises de dados de serviços de informações sobre patentes

Supervisão
Ary Mergulhão Filho

Coordenação
Jackson Max Furtunato Maia

Equipe Técnica interna
César Augusto Costa
Eduardo Amadeu Dutra Moresi
Ícaro Lorran Lopes Costa
Israel Garcia de Oliveira
Kleber de Barros Alcanfôr
Rogério da Silva Castro

Consultor
Jörg Neves Bliesener

Analista Administrativo
Larissa Martins Rocha

SUMÁRIO

1. INTRODUÇÃO

O reconhecimento e análise de informações existentes nas grandes massas de dados atualmente acessíveis permitem multiplicar a capacidade de atuação do CGEE, desde que técnicas adequadas de extração, tratamento e carga de dados sejam empregadas para reconhecer padrões que lhes sejam subjacentes. Nesse sentido, o projeto "Exploração de Dados e Visualização de Informações" (EDVI) visa fortalecer as competências do Centro, desenvolvendo e validando fundamentos, metodologias e ferramentas de análise de dados de CTI disponíveis, ampliando seu portfólio de serviços e auxiliando o embasamento metodológico das suas atividades.

Para consolidar no Centro metodologias para análises de dados de patentes, em particular, e demais tipos de propriedade intelectual no futuro, foram definidas metas de desenvolvimento de metodologias e ferramentas para análise de dados de patentes no contexto das atividades do projeto EDVI a partir de 2021. Essas iniciativas estão em consonância aos objetivos do CGEE, no sentido de que patentes são registros de etapas intermediárias do ciclo de maturidade tecnológica e são usualmente consideradas importantes indicadores de inovação. Espera-se que esses desenvolvimentos metodológicos e de ferramentas de análise auxiliem a internalização desse tipo de conhecimento no Centro e colaborem no embasamento em evidências dos estudos e demais trabalhos a serem realizados. Este texto relata os principais métodos de análise aprimorados ou desenvolvidos em 2022, iniciado por uma breve exposição sobre patentes.

2. PATENTES – ALGUMAS DEFINIÇÕES

De acordo com a Organização Mundial de Propriedade Intelectual – OMPI (WIPO, 2021), “Uma patente é um direito exclusivo, concedido para uma invenção, a qual é um produto ou processo que fornece uma nova forma de fazer alguma coisa ou oferece uma nova solução técnica para um problema. Para obter uma patente, a informação técnica sobre a invenção tem que ser divulgada para o público através de um documento de pedido de patente.”[1] Em caso de concessão, a patente garante ao seu dono o direito exclusivo de explorar comercialmente a invenção ou processo que foi criado. Esse direito é regulado em cada país através de leis que definem um órgão governamental para o gerenciamento desses pedidos e concessões. No Brasil, esse órgão é o Instituto Nacional da Propriedade Industrial (INPI)[2].

A estrutura do documento de submissão de uma patente pode variar entre países e escritórios, mas signatários do acordo internacional *Patent Cooperation Treaty* (PCT) seguem linhas gerais, para efeito de padronização. Essa padronização, particularmente a de metadados, facilita a elaboração de estratégias de análise, pois pedidos (ou concessões) de patentes de países diferentes têm a possibilidade de serem processados em conjunto. Como dados básicos, ressaltam-se: datas de publicação no país, código internacional de publicação, dados bibliográficos, como nomes de autores e de depositantes. O documento de patente é extremamente rico em informações, como reflexo da complexidade das relações comerciais e industriais relacionadas à propriedade intelectual e industrial. Com respeito aos métodos desenvolvidos no CGEE, as informações mais empregadas foram os resumos dos textos descritivos das inovações pretendidas na submissão de proposta de patentes e os códigos de classificação de campos tecnológicos envolvidos na invenção descrita.

[1] https://www.wipo.int/patents/en

[2] https://www.gov.br/inpi/pt-br

2.1 SISTEMAS DE CLASSIFICAÇÂO

Uma parte do processo complexo de submissão de um pedido de patente que merece atenção é a classificação da invenção em campos tecnológicos. Usualmente, a invenção não se limita a uma aplicação tecnológica e sua análise e classificação é feita caso a caso, o que requer do escritório de patentes a manutenção de um corpo especialistas de diversas áreas para realizar as classificações[3]. Com o objetivo de facilitar e padronizar o procedimento de classificação, foram criados os chamados sistemas de classificação. Existem vários sistemas, mas os mais importantes atualmente são o *International Patent Classification* (IPC) e o *Cooperative Patent Classification* (CPC). De acordo com a WIPO em (WIPO, 2021), o sistema IPC, estabelecido pelo tratado de Strasbourg em 1971 é atualmente utilizado por mais de cem países do mundo, fornece uma linguagem de símbolos organizada hierarquicamente para a classificação de patentes e modelos de utilidade de acordo com os campos tecnológicos aos quais eles pertencem.

A hierarquia do sistema IPC é dividida em cinco partes: seção, classe, subclasse, grupo e subgrupo. A seção tem o maior nível de granularidade dentro da hierarquia, representando uma classificação menos específica e o subgrupo tem o menor nível na hierarquia, representando uma classificação mais específica. De forma geral um código IPC é dado na forma XDDXD/DDDD, onde X representa letras e D dígitos numéricos.

Como exemplo, seja o código G06N3/063. Descendo na hierarquia temos G representando a seção “física” (que se refere genericamente a áreas de ciências exatas). G06 representando a classe “computação, cálculo e contagem. G06N representa a subclasse “sistemas de computador baseados em modelos computacionais específicos”. Descendo mais um nível, G06N3/00 representando o

[3] Com o avanço da inteligência artificial, será cada vez mais comum a realização de classificações automáticas de patentes, mas convém ressaltar que modelos de *machine learning* são treinados a partir de dados pré-classificados por humanos e os desempenhos de tais modelos é medido e monitorado.

grupo “sistemas de computador baseados em modelos biológicos” e G06N3/063 representa o subgrupo “usando meios eletrônicos” (em contraste ao “uso de meios analógicos”, que é representado pelo subgrupo G06N3/0635).

O sistema de classificação *Cooperative Patent Classification* (CPC)[4] pode ser considerado como um detalhamento maior do IPC, pois possui códigos adicionais no nível de subgrupo e também uma nova seção Y, onde são listadas patentes que envolvem tecnologias emergentes de aplicações transversais. Atualmente o CPC é mantido pelo Escritório Europeu de Patentes juntamente com o escritório de patentes do Estado Unidos (USPTO – *United States Patent and Trademark Office*).

Para trabalhos no CGEE que demandam a caracterização de tecnologias envolvidas na geração de inovação os sistemas de classificação de patentes são particularmente úteis. Cada sistema rotula campos tecnológicos com códigos padronizados pertinentes à invenção que se pretende proteger. A curadoria humana na classificação das patentes implica um mapeamento tecnológico acreditado e universal no nível de cada documento. Dessa forma, uma dada coleção de grandes volumes de dados de patentes contendo suas classificações permite uma caracterização tecnológica abrangente do conjunto que, em princípio, tende a ser mais precisa do que classificação de textos acadêmicos a partir de agrupamentos de palavras-chave (muito usado nos mapeamentos temáticos do CGEE). Isso é presumido porque, se para textos acadêmicos a classificação usualmente depende da percepção subjetiva dos autores sobre seus trabalhos, no caso das patentes os padrões delineados pelos sistemas de classificação mitigam a subjetividade. Além disso, no caso de patentes, em princípio não há limites no número de códigos atribuídos ao documento, enquanto que em artigos esse número normalmente é limitado.

[4] Descrições dos códigos de classificação CPC podem ser encontradas em todos os níveis da hierarquia, por exemplo, na página https://worldwide.espacenet.com/patent/cpc-browser#, com a separação entre códigos IPC e CPC.

2.2 BASES DE DADOS DE PATENTES

Independentemente do tipo de pesquisa, é necessário ter acesso a uma boa base de dados para encontrar patentes de interesse. Abaixo segue uma lista de bases às quais o CGEE tem acesso:

- Espacenet: desenvolvido pelo Escritório Europeu de Patentes, o Espacenet é um serviço web de acesso gratuito para busca e submissão de pedidos de patentes. O serviço apresenta uma interface de navegabilidade simples, embora limitada, com cerca de 100 milhões de patentes de vários países do mundo (não apenas da Europa). Sua base de dados é um dos recursos mais disponíveis para análises de patentes. O Espacenet tem como característica marcante o suporte para pesquisas usando o sistema CPC de classificação. A página pode ser acessada pelo endereço https://worldwide.espacenet.com.

- Escritório de patentes do Estados Unidos: o principal banco de dados de patentes dos Estados Unidos, o USPTO permite pesquisa de patentes que datam de até 1790. Apesar de possuir uma interface pouco amigável, esse banco de dados possui uma boa documentação para criar expressões de busca mais sofisticados. A página pode ser acessada pelo endereço https://patft.uspto.gov/netahtml/PTO/search-bool.html.

- Patentscope: base de dados mantida pela WIPO, o serviço possui uma interface bastante amigável, mesmo para buscas mais sofisticadas. Apesar de não possuir suporte para pesquisas utilizando o sistema CPC, o Patentscope é efetivo para buscas utilizando o sistema IPC. Pode ser acessado em https://patentscope.wipo.int/search/en/search.jsf.

- Lens.org: base de dados australiana, o Lens (anteriormente "*Patent Lens*") é uma das opções mais completas disponíveis atualmente para aplicações não comerciais (embora tenha taxas de licenciamento para alguns casos) e possui suporte aos sistemas IPC e CPC. Com sua interface bastante amigável, é possível rapidamente realizar pesquisas complexas. O serviço também permite

visualizações dos resultados das pesquisas, como quantidade de patentes por código, códigos IPC ou CPC mais frequentes e entidades que mais depositam patentes na determinada área. O Lens pode ser acessado em https://www.lens.org.

• Derwent: Plataforma disponibilizada pela *Clarivate Analytics*, mas acessável via Portal Periódicos da CAPES com algumas restrições de volume de downloads, a Derwent contém patentes de mais de 40 países. Uma das suas maiores vantagens é a existência de curadoria própria na classificação de documentos, o que facilita a construção de mapeamentos tecnológicos. A base opera apenas com classificação IPC.

• Scopus: com as mesmas características da Derwent, o serviço de patentes da Scopus também pode ser acessado pelo Portal Periódicos e é disponibilizado pela empresa Elsevier.

• INPI: A partir de dados não estruturados da revista do INPI, o CGEE constituiu uma base de dados de patentes brasileiras. Nessa base há metadados cadastrais que normalmente não existem nas bases listadas acima. A possibilidade de realizar cruzamentos entre os dados para enriquecê-los abre boas oportunidades de análise para o Centro.

• Base Lattes: Apesar de pouco utilizada para esse fim, a base de currículos Lattes tem espaço para a produção tecnológica de seus usuários, incluindo patentes. Apesar do conteúdo ser bastante limitado com relação às bases de dados de patentes, o cruzamento de identificadores únicos de patentes com os identificadores únicos de pesquisadores do Lattes permite a realização de meta-análises de produção acadêmica associada à produção tecnológica de pesquisadores.

3. ATIVIDADES DE DESENVOVIMENTO DE METODOLOGIAS EM 2022

Em conformidade ao fluxo de trabalho do projeto EDVI, o desenvolvimento de protótipos funcionais é precedido por soluções para problemas relacionados a dados trazidos pelos demais projetos do Centro. O processo envolve o esclarecimento das perguntas norteadoras, definição de bases de dados que podem ser analisadas para a obtenção de evidências para as respostas, a aquisição das bases de dados, limpeza e normalização desses dados, análise exploratória, definição de algoritmos adequados, mineração de padrões e preparação das respostas por meio de relatórios, visualizações ou, quando o mesmos tipo de resposta pode ser de interesse de outros projetos, a implementação dos algoritmos desenvolvidos em protótipos funcionais em software. Esse processo, dependendo do problema, tem como produtos soluções em diferentes níveis de maturidade tecnológica. Nesta seção, descreveremos brevemente as atividades que precederam as soluções que avançaram em nível de maturidade:

a) Prova de conceito sobre a exploração de patentes a partir de uma lista de empresas/CNPJ com cruzamento de dados de patentes da base extraída da coleção de Revistas da Propriedade Industrial, editadas pelo INPI, e a base "CNPJ Dados Públicos" da Receita Federal. Os resultados foram entregues ao projeto "Serviço de assessoramento no monitoramento, avaliação e produção de subsídios técnicos para a inovação". Esse trabalho foi muito importante para o entendimento dos dados do INPI e para a identificação de lacunas e oportunidades de melhoria da nossa base de patentes nacionais, que hoje contempla mais de 900 mil registros. A atividade motivou também a realização de testes mais realistas da base de dados de patentes brasileiras e testes de validação comparando os dados do Brasil na base, na base Derwent e na base Lens.org.

b) Início de estudos e testes de métodos para identificação de emergência tecnológica a partir de análises de dados de coocorrência de códigos IPC, com

base em proposta da RAND Corporation. O método original utilizava o sistema de códigos antigo do USPTO e focava na evolução do grau médio (coocorrências entendidas como arestas entre dois códigos). Os estudos, ainda em andamento, testam a hipótese de melhoria de resultados com análises da evolução do grau ponderado de códigos IPC, além de considerarem diferentes classes de granularidade.

c) Prova de conceito do uso de métodos de detecção de anomalias para o estudo de emergência tecnológica.

d) Prova de conceito de detecção de emergência tecnológica usando métodos baseados no *embedding Glove* e redes neurais recorrentes;

e) Elaboração de uma biblioteca em Python especializada na análise de dados de patentes de diversas fontes, inicialmente preparada para tratar dados nos formatos da Derwent e da Lens.org.

f) Lançamento de ferramenta de visualização rápida das distribuições de frequências de códigos IPC de conjuntos de dados de patentes no formato Derwent. Este protótipo permite a visualização interativa do espectro de códigos IPC/CPC de patentes em todos os seus níveis de granularidade. O Software consiste em uma interface web baseada na tecnologia NuxtJS (JavaScript) e também um backend criado usando a tecnologia FastAPI (Python). A ferramenta decorrente desse trabalho tem sido utilizada e validada nas análises realizadas pela equipe do projeto "Agenda positiva da mudança do clima e do desenvolvimento sustentável".

g) Participação, junto com a equipe do projeto “Agenda positiva...” na elaboração de proposta de Acordo de Cooperação com o INPI. O conjunto de resultados exposto acima capacitou o Centro a estabelecer uma colaboração profícua com INPI que certamente permitirá melhorias seja nos nossos métodos, seja nos nossos softwares.