Como a ciência se transforma em tecnologia no Brasil por meio da Análise de indicadores de Patentes

Resumo

Abstract

Introdução

Este presente trabalho tem por objetivo avaliar quais áreas da ciência são mais relevantes na transição tecnológica no Brasil. Esta análise está inclusa no escopo de análise bibliométrica de patentes e ‘*tech mining’*, área que está progressivamente evoluindo no cenário de *text mining* nas últimas décadas (Madani, 2015).

A partir da análise e do mapeamento da conversão da ciência em tecnologia, é possível, por exemplo, a setores estratégicos de companhias avaliar quais são os caminhos e embasamento teórico científico necessários para o desenvolvimento de determinado produto ou invento tecnológico (Madani; Weber, 2016) ou, em âmbito macro-econômico, estimar o perfil futuro da ciência e da tecnologia no país com base na evolução consequente de áreas correlatas (Lahm, 2018).

Os indicadores para avaliação da transição entre ciência e tecnologia neste estudo consistem na análise da citação reversa de determinado invento ou metodologia, quando um documento de patente cita determinado artigo científico de área previamente estabelecida (Glanzel; Meyer, 2003). Portanto, para esse estudo, foram necessários o emprego de um conjunto de técnicas de Ciência de Dados para a extração dos documentos de Patentes, como também a categorização automática da área cientifica de um grande volume de dados de documentos.

A partir da categorização e do mapeamento das áreas da ciência, pretendemos representar os dados obtidos em formato de Ontologias, pois, dessa forma, ampliamos a governança de dados em torno dos dados de patentes, assim como para qualquer outro dado representado em formato de *Linked Data* (Gartner, 2019). Também aproveitamos, por meio da representação do conhecimento obtido em vocabulário de ontologias, OWL e RDFS, o emprego de técnicas de raciocínio automático para estabelecimento de novos relacionamentos implicitos no nosso conjunto de dados.

“Raciocinadores automatizados podem deduzir (inferir) conclusões do conhecimento fornecido, tornando explícito o conhecimento implícito” (Antoniu; Groth; Harmelen; Hoekstra, 2012, p.13).

Dessa forma, promovemos o trabalho colabarativo em ciência de dados e patentes, pois tornamos mais simples o acesso a informação e ao conhecimento de Patentes, exigindo somente o conhecimento dos termos do vocabulário ontológico de patentes para se ter acesso aos dados.

Metodologia

Métodos de Categorização de área científica

A classificação dos domínios de áreas cientifica de patentes obedece a classificação proposta pela empresa de avaliação científica science-metrix, das quais estão presentes as seguintes áreas:

|  |  |
| --- | --- |
| Área da Ciência |  |
| Agriculture, Fisheries & Forestry |  |
| Built Environment & Design |  |
| Enabling & Strategic Technologies |  |
| Engineering |  |
| Information & Communication Technologies |  |
| Communication & Textual Studies |  |
| Historical Studies |  |
| Philosophy & Theology |  |
| Visual & Performing Arts |  |
| Economics & Business |  |
| Social Sciences |  |
| General Arts, Humanities & Social Sciences |  |
| General Science & Technology |  |
| Biomedical Research |  |
| Clinical Medicine |  |
| Psychology & Cognitive Sciences |  |
| Public Health & Health Services |  |
| Biology |  |
| Chemistry |  |
| Earth & Environmental Sciences |  |
| Mathematics & Statistics |  |
| Physics & Astronomy |  |
|  |  |

Fonte:Science-Metrix

Para realizar a classificação da área científica de uma patente utilizamos duas vias:

* Análise de Citação Reversa

A classificação categoriza cerca de 15000 revistas ciéntificas de acordo com a área mais relevante em suas publicações.

A categorização de *Journal* ou revista científica, para nosso proposito, serve como recurso para a aplicação da técnica de avaliação de citação reversa, pois os documentos de patentes, em seu corpo, podem apresentar citações a artigos científicos publicados nesses Journals. A partir de um casamento de uma citação de uma patentes com um Journal, atribuímos a correspondente área científica do *Journal* à patente.

* Atribuição de Área Científica por meio de identificação de termos exclusivos do domínio textual

Conforme proposto por Custódio (2019, p.46) os diferentes textos apresentam em comum termos referentes a um domínio textual, e esses vocabulários, quando exclusivos de determinado domínio, apresentam cárater discriminante da área científica se identificados em determinado texto.

Portanto, com base na análise de um modelo supervisionado de Bag of Words, que, basicamente, consiste na construção de um histograma de palavras naquele determinado conjunto de textos. (Zhang;Jin;Zhou, 2010) <https://link.springer.com/article/10.1007/s13042-010-0001-0> mostram que o modelo de Bag Of Words é um modelo muito relevante e preciso na Categorização de domínios de PLN, pois ele é um modelo genêrico e simples para a construção de diferentes conjuntos de dados e domínios. Isso implica que não necessitamos realizar a definição de heurísticas específicas para a categorização de cada domínio.

(Matsubara;Martins;Monard, 2003) <https://www.researchgate.net/profile/Maria-Carolina_Monard/publication/266865610_PreTexT_uma_ferramenta_para_pre-processamento_de_textos_utilizando_a_abordagem_bag-of-words/links/544e6ac30cf26dda089010d2/PreTexT-uma-ferramenta-para-pre-processamento-de-textos-utilizando-a-abordagem-bag-of-words.pdf> por outro lado, abordam os problemas de alta dimensionalidade apresentada pela técnica de Bag Of Words, da qual cada nova palavra identificada no conjunto de dados torna-se uma nova componente no modelo de classificação. Por isso, são explanadas uma série de técnicas para redução de dimensionalidade e expressarmos somente os termos mais relevantes no conjunto de Bag Of Words.

Contudo, nosso modelo prioriza a categorização específica dos domínios de áreas, portanto pretendemos representar todas as componentes na técnica de Bag of Words, com o intuíto de não perder variância explicada pelo modelo, além de priozirar a identificação de termos técnicos e exclusivos de determinada área científica.

Nosso modelo de Bag Of Words atual cobre um conjunto de 123 mil palavras em inglês, do qual estima-se que cobre algo entre 10% e 25% das palavras do idioma inglês.

Abaixo, apresentamos a representação de uma *Cloud of Words*, técnica visual equivalente a estrutura de dados *Bag Of Words*:

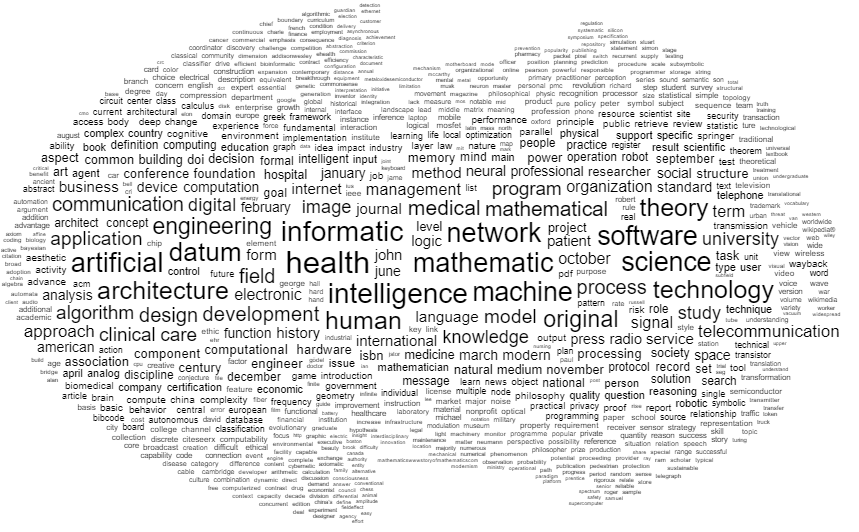


Imagem 1:Termos mais relevantas nas áreas de informação e comunicação tecnológica. Fonte: Arquivo Próprio

Para cada uma das áreas, foram construídos modelos semelhantes contendo conjuntos de palavras relevantes para aquelas determinadas áreas.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

