特許文書情報を用いた発明内容の抽出と出願人タイプ別特性比較

文部科学省 科学技術・学術政策研究所
第 2 調査研究グループ

要旨

本稿では，特許の発明内容を分析するための自然言語処理技術と統計数理手法に基づく新たな手法を提案し，日本の特許データを用いて提案手法の機能可能性を評価した。結果として，特許の発明内容の分布状況の可視化や類似特許の検索において提案手法が期待通りに機能することが確認された。また，本提案手法により，日本では個人や大学等の特許は幅広い分野に分布している一方，企業特許は特定分野に集中的に出願されていることが分かった。

研究開発に関する情報は企業にとって戦略的に重要なものであり，内部情報として企業の内部で秘匿されることが多い。しかし，特許が出願されると，その発明の内容は広く公開される。そのため，特許データは個々の企業や産業，場合によっては国全体の技術トレンドについて分析するための貴重な情報源となっている。また，特許権の構成要件として，当該発明の新規性や進歩性に加えて，産業応用可能性が必要とされる。そのため，科学技術論文として公開される情報と比べて，特許情報には，より産業寄り，言い換えれば新商品などのイノベーションに近い情報が含まれている。

他方，特許の情報はデータサイズが膨大になるため，単純にその内容の類似度で分類することは計算コストの面から難易度が高かった。これらの課題に対応するため，本稿では分散表現などの近年普及してきた自然言語処理手法及び高次元ベクトル近傍探索，次元圧縮などの統計数理手法を用いた特許データの分析を試みた。まず，日本の特許庁の公開公報情報におけるタイトルと要約文を用いた分散表現を通じて，特許内容のベクトル空間モデルを作成した。次に，この特許内容のベクトル空間モデルを用いて，特許のクラスタリングや近傍特許の抽出，特許間の距離の測定を試行した。さらに，これらの情報を用いて出願人タイプ（個人・企業・大学等）による特許の特性を明らかにした。

A method of extracting content information from patent documents and comparison of their characteristics by applicant type by using the vector space model of distributed expressions

2nd Policy-Oriented Research Group,
National Institute of Science and Technology Policy (NISTEP),
MEXT

ABSTRACT

In this paper, we propose a new method based on the latest natural language processing technology and statistical mathematical methods for analyzing patent invention contents, and evaluate the usefulness of the proposed method using Japanese patent data. As a result, the usefulness of the proposed method was confirmed in the visualization of the distribution of the invention contents of patents and the search for similar patents. In addition, the proposed method shows that patents by individuals and universities are distributed in a wide range of fields in Japan, while company patents are intensively applied in specific fields.

Information related to research and development is strategically important for companies, and is often hidden inside the company as internal information. However, when a patent application is filed, the contents of the invention are widely disclosed. For this reason, patent data is a valuable source of information for analyzing technology trends in individual companies, industries and, in some cases, the entire country. In addition to the novelty and inventive step of the invention, industrial applicability is required as a constituent of patent rights. Therefore, compared to information published as scientific and technical papers, patent information contains information that is closer to industry, in other words, closer to innovation such as new products.

On the other hand, since the data size of patents is enormous, it is difficult to simply classify based on the similarity of the contents in terms of calculation cost. In order to deal with these problems, this paper tried to analyze patent data by using natural language processing techniques such as distributed expressions and statistical mathematical techniques such as high-dimensional vector neighborhood search and dimension compression. First, a vector space model of patent contents was created through distributed representations using titles and abstract sentences in the publication information of the Japanese Patent Office. Next, using the vector space model of this patent content, we tried clustering patents, extracting neighboring patents, and measuring the distances between patents. Furthermore, the characteristics of patents by applicant type (individual, company, university, etc.) were clarified using this information.