2P2-9

自然言語処理を用いた特許文書の可視化手法

A Visualization Method of Patent Documents using Natural Language Processing

中沢 誠[*1]
Makoto Nakazawa

吉田 博[*1,*2]
Hiroshi Katayama-Yoshida

[*1] 大阪大学産業科学研究所
ISIR, Osaka University

[*2] 大阪大学大学院基礎工学研究科
Graduate School of Engineering Science, Osaka University

We develop a new visualization method of patent documents in a two-dimensional plane. By using morphological analysis, only nouns are extracted from claim, and then these nouns are plotted using eigenvectors of co-occurrence matrix. The proposed method enables us to get a visually-apparent map representing technical content of the patent document accurately.

1. はじめに

近年、産学連携あるいは企業間の技術分析の場面において、自他の技術の相違を理解しやすい形で提示する手法ー特に可視化手法ーの開発が求められている。その際解析の対象とされるのは、情報共有が行い易いという点から、多くの場合特許文書(特許明細書)である。

特許文書を解析する場合、通常、自然言語処理における形態素解析により特許請求項を形態素に分解する。そして多くの場合、形態素のうち名詞のみを取り出し、それら名詞がどの特許文書に何回現れるか(出現頻度)に着目し、名詞と特許文書との関係について議論している。ただし、この手法では、同一の特許文書に現れる名詞同士の関係が取り込まれないため、解析後に表現される技術内容が不明確となってしまう。

本研究は、この問題点に鑑み、特許請求項の構造に着目し、複数の特許文書から構成される共起行列を用いることで、特許文書に記載されている技術内容を、明確にかつ正確に把握する手法の開発を試みるものである。

2. 関連研究

特許文書を俯瞰分析するツールとしては、「テクノロジー・ヒートマップ分析」が知られている[三宅 04]。このツールでは、主成分分析により特許文書と技術用語との関係が解析され、文書の密度分布が色の濃淡で二次元平面上に表現される。また、語の共起に注目し、文書からキーワードを抽出する研究としては、例えば[松尾 02] や[大澤 99] が知られている。本研究は、語の共起に注目するが、単一のキーワードを抽出するのではなく、用いられている複数の名詞を抽出し、文書に記載されている技術内容を正確に表現可能とする点に特徴がある。

3. 本研究のあらまし

3.1 特許請求項の構造

特許請求項の記載形式としては、主にジェプソン形式(「・・・において、・・・を特徴とする」)、あるいは要件列挙形式(「〜と、〜と、〜と、を備える・・・」)が用いられている。例えば後者の形式において、「A, B, C の中から選ばれる D と、a, b, c の中から選ばれる d と、を備えた X」と記載されている場合、A, B, C 及び D (a, b, c 及び d) は関係が深く、当該文章において『共起』していると見なせる。また、D と d は X の構成要素であり、これらは強いつながりを有している(つまり、D, d 及び X も広義には『共起』していると見なせる)ことがわかる。

そこで、名詞をノードとし、共起している名詞間にリンクを張ると、請求項を一種のネットワークと捉えることができる。そして、当該ネットワークにおいては、技術的に関係が深い名詞である A, B, C 及び D (a, b, c 及び d) はクラスターを形成し、さらに D 及び d という構成要素は大きな媒介中心値(中心性を表す指標の一つ)を有することがわかる。このネットワークを行列の形で表現するには、ある一つの特許文書(p)中で名詞 i と名詞 j とが共起している回数(共起頻度)w $^{p}_{ij}$ を、i 行 j 列における行列要素とすればよい。複数文書の場合には、例えば共起頻度の和を対応する行列要素とすることが考えられる。

3.2 名詞及び特許文書の二次元平面への配置

このように請求項の構造に着目して作成された共起行列を対角化し、例えば最大及び二番目に大きい固有値に対する固有ベクトルを、それぞれ平面における各名詞の x, y 座標とする。これにより、請求項中で共起している名詞は平面上で近い位置に配置される。つまり、技術内容を構成する複数の要素(例えば前述の A, B, C 及び D)は平面上でクラスターを形成し、関連が弱い名詞は離れた位置に配置される。

また、名詞が配置されたのと同じ平面に文書をプロットする場合、名詞と関連付けられた一点に文書の座標を特定してしまうと、当該文書に記載されている技術内容と(当該技術内容を構成している)名詞との位置関係が不明確になる虞がある。そこで、本研究においては、ある文書を表現する場合、当該文書に現れるすべての名詞を用い、それら名詞の各座標位置に当該名詞の(その文書における)出現頻度に相当する重み(複数文書の場合には、それらの和)が与えられているとして扱う。これにより、特許文書に記載されている技術内容を、明確にかつ正確に把握することが可能となる。

参考文献

[三宅 04] 三宅 将之, 宗 裕二, 姫野 桂一: 戦略的知財ポートフォリオ・マネジメント, pp. 4-17, 知的資産創造, 2004 年 10 月号.

[松尾 02] 松尾 豊, 石塚 満: 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム, 人工知能学会論文誌, Vol.17, No.3, pp. 217-223 (2002).

[大澤 99] 大澤 幸生, Nels Eric Benson, 谷内田 正彦: KeyGraph : 語の共起グラフの分割・統合によるキーワード抽出, 電子情報通信学会論文誌, D-I, Vol. J-82-D-I, No. 2, pp. 391-400 (1999).

連絡先: 中沢 誠, 大阪大学産業科学研究所, 大阪府茨木市美穂ヶ丘8−1, nakazawa@cmp.sanken.osaka-u.ac.jp