

○中村達生（三菱総研），玉田俊平太（筑波大先端学際領域研）

## 1 研究目的

今般、研究開発により、我が国経済の国際競争力を確保することは、科学技術政策の大きな目標の一つであり、我が国経済における「技術」の果たす役割の重要性がより一層たかまりつつある。それには技術が研究段階から、産業技術、製品、社会へと波及する流れを客観的な定量データで示す事が必要であり、本研究では、データマイニング(概念検索)手法を用いることで、これらの技術関連をマクロ的に分析し、手法の有効性と適用の可能性を示すことを目的としている。

## 2 従来の研究

技術関連を表す考え方には、論文件数や特許件数を指標として用いる方法と、特許のサイテーション（引用）情報を用いる方法が存在する。前者の方法は技術の関連は示せるが、親と子の関係、すなわち、どちらが引用元であるかは判りにくい。ただし、比較的容易に全分野にわたる関連度(件数)を把握することが可能である。一方、後者のサイテーションは、個々の引用情報をひもといて分析するため、全分野を網羅する分析は難しいが、個別技術毎に技術の流れを明示することができる。ただし、米国特許では制度的に、特許の明細書中に参考文献として引用した論文や特許のタイトル、文献名等の情報が記載されることになっているため、比較的広い分野にわたって分析することが可能である。これらの引用文献のうち、特許一件あたりの科学論文の件数を集計したのがサイエンスリンケージ(Science Lincage)である[1]。特許における論文の引用は、技術(特許)とそれが依拠する科学とを関係づけるものと考えられ、したがって、その件数は科学との関連性の強さを示すと解釈できる。さらに、特許の出願者による引用ではなく審査官による引用であるため、比較的客観性が高いとされている。

しかし、日本国特許においては論文引用情報を記載する制度がないため、参考とした論文書誌情報の記述が極めて少なく、多くの場合は、米国特許を用いてサイエンスリンケージの分析を行うことになる。ところが、この方法では米国特許に出願していることが前提となるため、我が国の産業技術(特許)をすべて網羅していることにはならない。また日本国内から分析する場合は、米国特許の検索システムの仕様による分析上の制約が存在する。そこで、本研究では、日本国特許データの内容的な類似性から関連を表す、概念検索を用いた分析手法を提案する。

## 3 概念検索を用いた技術関連分析の方法

### 3. 1 概念検索を用いた技術関連分析とは何か

概念検索を応用した技術関連分析手法とは、対象とする技術テーマの概要を入力文として特許の全内容を検索し、その類似性を定量指標で表すものである。この方法では、類似特許の明細書の全内容を検索するため、漏れのない検索が可能であり、また、従来の分野・分類にとらわれずに抽出し、類似性を定量的に表すことが可能である。

### 3. 2 概念検索の仕組み

概念検索では、あらかじめデータベース中の各文章をベクトルで表現しておき、入力した文章(技術テーマの概要)のベクトルと方向が近いものほど、内容が類似しているはずであるとして、抽出と序列化を行っている[2]。文章をベクトル化するには、形態素解析と

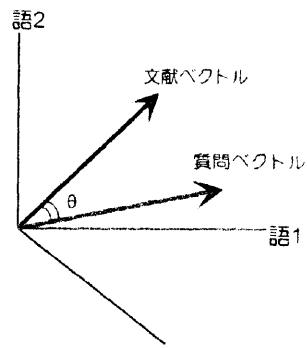


図 1 ベクトル空間モデル (三次元の例)

呼ばれる方法を用いて複数の単語に分割し、各単語の重要性は、文章とデータベース中に現れる頻度から決定する。形態素解析とは、一般には、文を辞書に登録されている語へと分解する処理を意味し、大きく分けて、①語の切り出し、②接辞処理の2つの段階からなる。切り出された単語への重み付けは、理論的には次のようにして決定されている(式1参照)。ある単語が文献の中に繰り返し出現する頻度(TF)が高く、かつ、その語を含む文献が一部に偏って出現(IDF)している場合には、その単語は重要であると判断する。機能語や一般語の場合、前者の条件(TF)だけが高くなるので、自然に除外される。一つの文献や検索に用いる文章(質問ベクトル)は、これらの重み付けされた要素ベクトルの合成ベクトルで示すことでできる。文献ベクトルと質問ベクトルの方向が近いほど類似性が高く(図1)、ベクトルの近さは類似度と呼ばれる内積を用いた指標であらわされる。

$$W_{ij} = TF_{ij} \times IDF_j \quad \dots \text{式1}$$

TF<sub>ij</sub> : Term Frequencyの略。文献iの中に出現する語jの頻度

IDF<sub>j</sub> : 逆文献頻度(Inverse Document Frequency)の略。語jを含む文献数の逆数。

### 3. 3本研究における概念検索手法の適用方法

#### (1) 対象データベース

本研究では、いくつかの重要技術分野に関する論文と産業分野を対象として、それらの定義を入力文として検索に供した。

特許情報は、電子ファイル化されて分析に供することができる全情報、すなわち H15 年以降に登録された全特許(93 年~2000 年 9 月登録分)を対象とした。

#### (2) 検索文の入力から類似特許の検索までの流れ

はじめに特許データ(および辞書データ)情報をデータベースに登録し、インデキシング(形態素解析、重み付け)を実施し、あらかじめベクトル化する。つづいて、対象とする技術分野と産業分野の定義文を検索文として検索を実行する。最後に、抽出された特許の書誌情報に基づいて、時系列分析、関連技術分野の分析を行う。

## 4 分析結果

概念検索を用いると、従来のカテゴリーやキーワードにとらわれずに技術関連分析が可能であり、意外な分野から類似技術を発見することもできる。論文発行年や特許出願年に着目して時系列分析を実施すると、技術分野毎のタイムラグ、類似する技術の変遷、さらには市場規模との相関を知ることができる。

### 4. 1 技術と技術の相関

例として製版業に類似する特許を抽出し、明細書に記載されている IPC コード(国際特許分類)を WIPO 分類に従って件数を整理すると、Unit9(印刷、筆記具、装飾)と Unit26(測定、光学、写真、複写機)に関する分野の件数が最も高くなった(図2)。いずれも製版業にまつわる技術分野を正しく抽出しており、概念検索を用いたことにより、異なる分野からも類似する技術を抽出できた事例と言える。

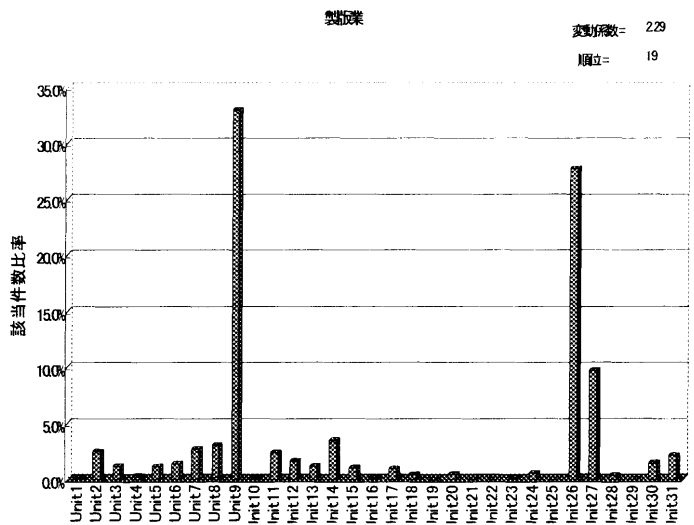


図2製版業の関連技術分野(WIPO 分類)



