検索精度向上への取り組み
類似文献検索の特許検索への適用に係る検討2

一般財団法人工業所有権協力センター　研究所総括研究員　土居　仁士

PROFILE
平成21年10月より現職

1 はじめに

一般財団法人工業所有権協力センター（IPCC: Industrial Property Cooperation Center、以下「財団」と表す。）研究所は、財団の主たる事業である検索事業、一元付与事業のさらなる効率化をめざし、IPCCシソーラス等の独自データ資産を整備するとともに、それらの一層の活用手法を検討している。

そして、2年前から、検索事業への適用の可能性を探るため、IPCCシソーラスを適用して類似文献検索を行うことができるシステムのプロトタイプシステムを構築し、その評価を行っている。

2 プロトタイプシステム

プロトタイプシステムは、情報処理振興事業協会（IPA）が実施した独創的情報技術育成事業の研究成果である汎用連想計算エンジン（GETA）を用い、日本語形態素解析器としてChaSenを、形態素解析用辞書としてIPADICを利用するとともに、抽出された特徴語にIPCCシソーラスを適用して類義語を追加する機能を追加したものである。

なお、検索対象となる母集合は、1994年から2007年に公開された公開特許公報約520万件で構成している。

3 形態素解析の改良

このプロトタイプシステムを用いた類似文献検索は、昨年報告したとおり、IPADICによる形態素解析のままでは、適切な技術用語を特徴語として抽出できないばかりでなく、IPCCシソーラスの代表語に多く含まれる複合語が一般的な単語に分解されてしまい、複合語の類義語拡張が適切に行えないという課題を有していた[1]。

例えば、電気分野で用いられる「継電」という単語が、「継」と「電」に分解されて特徴語として抽出される等の現象が発生するなど、分解された各単語の意味が本来の特徴語の意味と異なることとなり、結果として、意図しない文献が上位に抽出される可能性を有していた。

この課題に対して、以下の方式を用いて複合特徴語を補完し検証を行った。

複合名詞抽出方式

Query文や検索対象文書のテキスト中の連続する名詞や未知語を連結して作成した複合名詞を複合特徴語として抽出し、基本特徴語とは別に設定した重み係数で利用する方式

関連複合語抽出方式

IPCCシソーラスに蓄積された代表語のうち、Query文や検索対象文書のテキスト中に出現するものを複合特徴語として抽出し、基本特徴語とは別に設定した重み係数で利用する方式

ハイブリッド抽出方式

複合名詞抽出方式と関連複合語抽出方式を組み合わせた方式

4 検証方法

基本特徴語と上記各抽出方式にて抽出された複合特徴語に対し、IPCC シソーラスによる類義語の拡張を行い、上位 100 件での再現率と平均精度の検証を行った。

複合名詞抽出方式については、部分複合特徴語のみの抽出、最長複合特徴語のみの抽出、その両者の抽出を行う方法の検証を行った。

ハイブリッド抽出方式においては、４つの抽出方式で共通して抽出された複合特徴語を採用する方法、２つの抽出方式のいずれかで抽出された複合特徴語を採用する方法の検証を行った。

また、複合特徴語の重みについては、基本特徴語の重みの 10%、50%、100%、1000%の４パターンにて検証を行った。

5 検証結果

複合名詞抽出方式については、部分複合特徴語と最長複合特徴語の両者を複合特徴語として利用し、複合特徴語の重みを 50%とした場合が最も高い再現率・平均精度が得られた。

関連複合語抽出方式においては、複合特徴語の重みを 10%とした場合が最も高い再現率･平均精度が得られた。

ハイブリッド抽出方式においては、２つの抽出方式のいずれかで抽出された複合特徴語を利用し、複合特徴語の重みを 10%とした場合が最も高い再現率･平均精度が得られた。

そして、各抽出方式における再現率・平均精度について複合特徴語を利用しないベースラインと比較したところ、

ベースライン　＜　複合名詞抽出方式

複合名詞抽出方式　＜　関連複合語抽出方式

関連複合語抽出方式　＜　ハイブリッド抽出方式

となり、複合特徴語を利用することの有効性、特に、IPCC シソーラスから複合特徴語を抽出することの有効性が確認された。

図 1 のグラフは、ハイブリッド抽出方式（最適方式）と、ベースラインの再現率をテーマコード毎に比較したものであり、幅広い分野での改善が認められた。

6 おわりに

類似文献検索における IPCC シソーラスの活用については、一定の有効性が確認されたものの、今回の検証は、公開特許公報の全文を検索対象とし、Query 文として、要約＋請求項１を用いるなど、必ずしも検索業務の実務に即した検証とはなっていない。また、プロトタイプシステムは、静的な WAM を利用しており、既存の FI/Fterm 検索、全文テキスト検索との融合も十分に考慮されていない。

今後は、より実務に即した形態での類似文献検索のあり方の検討を進め、プロトタイプシステムの改良とともに実務への適用に向けて検証を進めていく予定である。

[参考文献]

[1] 居島一仁，検索精度向上への取り組み（類似文献検索の特許検索への適用に係る検討），Japio 2009 YEAR BOOK，pp. 168-171，2009.

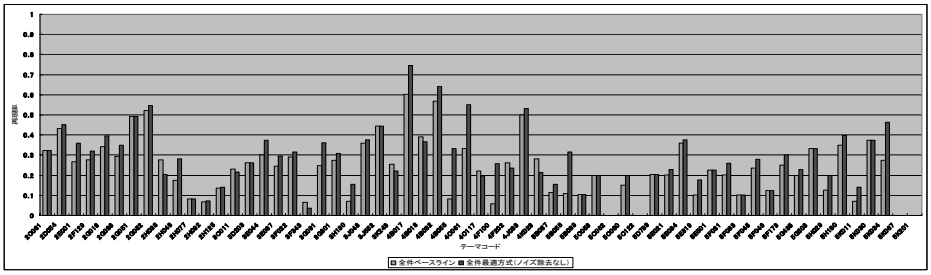

図１　再現率の比較