Japio

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

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
居島　一仁

PROFILE
平成 19 年より現職

1 はじめに

飛躍的に増加する検索対象特許文献数及び日々高度化する技術に伴う検索負担の増加への対応や年々増大が予想される検索業務量への対応は、現在、一般財団法人工業所有権協力センター（以下、「財団」と表す。）が解決すべき喫緊の課題となっている。

一方、財団は、IPCC シソーラス等の財団独自のデータ資産を有しており、これらを効果的に検索業務や一元付与業務に活用する手法を検討することにより、上記データ資産を活用した事業の更なる効率化ができるものと考える。

そのため、財団が実施している特許文献の検索事業で用いている現行検索システムと組み合わせて適用・評価できる類似文献検索システムに対し、如何にして IPCC シソーラス等財団のデータ資産を活用しうるかという観点から、類似文献検索システムの基本的な評価等の調査研究を行った。

2 評価用試験環境

本調査研究の類似文献検索システムとして、情報処理推進機構（IPA）が実施した独創的情報技術育成事業の成果である汎用連想計算エンジン GETA（Generic Engine for Transposable Association）を用い、形態素解析に用いる辞書として IPA が策定した IPA 辞書を採用し、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プロトタイプを作成した。

また、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 500 万件で構成した。

3 調査方法

検索報告書を作成した特許出願をテスト本願として、物理、機械、化学及び電気の各技術分野から３～４テーマ、計 14 テーマ、141 件のテスト本願を選定した。

各テスト本願において検索報告書に記載された X 文献（本願発明と引用発明の構成が同一であり、単独で引用可能な文献）、Y 文献（本願発明と構成が一部相違するものの、他の文献との組み合わせにより本願発明に容易に想到し得る文献）を正解文献とし、それらが、類似文献検索システムのプロトタイプで上位 100 件以内の第何位にランクされるかについて調査を実施した。

この結果から、再現率等を算出し、IPCC シソーラス適用の有無による比較等を行った。

・再現率：検索結果の上位 N 件の中に、正解文献がど

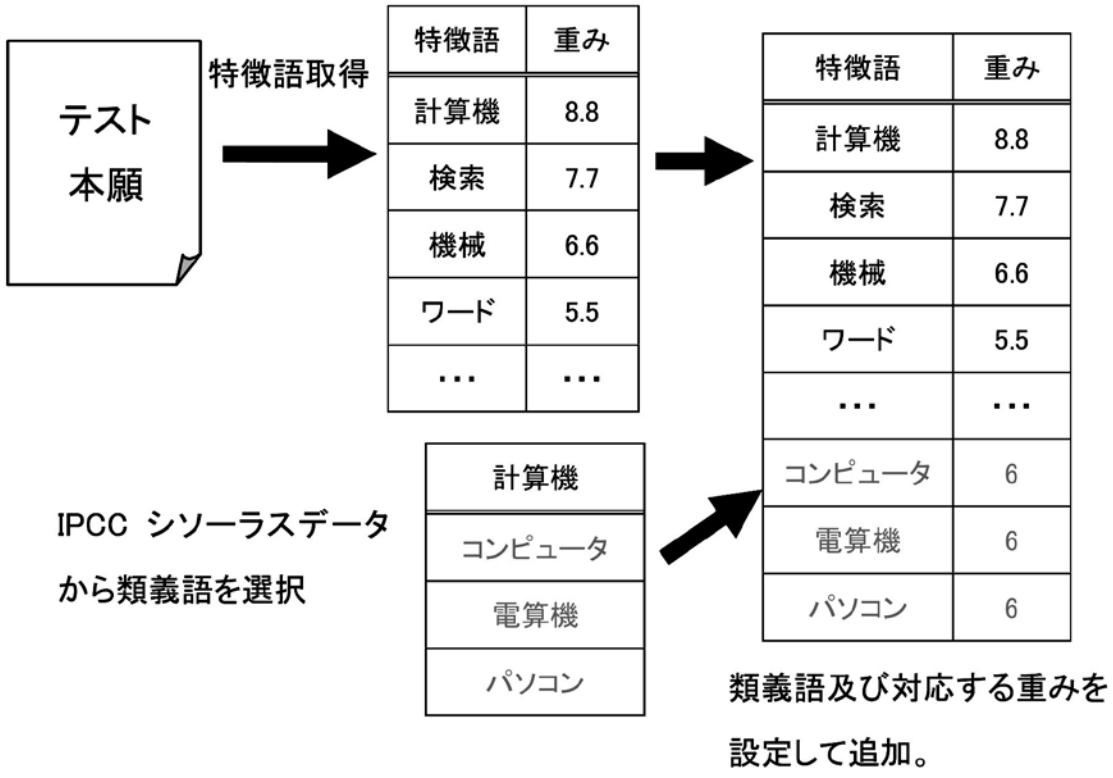

図１　特徴語への IPCC シソーラスデータの適用

のくらいの件数含まれているかの割合。

（再現率）＝（上位 N 件の中に含まれる正解文献数）
÷（正解文献数）　・・・・・（式１）

4 検索条件

検索は、所定の検索条件を設定して一括的に処理する方法（バッチ検証）と、評価者がテスト本願を１件ずつ検索する方法（オンライン検証）の２種類を実施したが、以下、オンライン検証での検索結果について述べる。

類似文献検索時 IPCC シソーラスの適用は、評価者が、テスト本願を理解した上で、テスト本願から抽出された特徴語から、テスト本願が属するテーマに属する IPCC シソーラス内に記載されたその特徴語の類義語を選択の上、各類義語に対し、元の特徴語の重みを参考に適切な大きさに設定した上で、特徴語として追加したものをテスト本願の特徴語群として類似文献検索を行った。

特徴語抽出は、テスト本願及び検索対象となる公開特許公報ともに全文から行う条件で実施した。

5 調査結果及び今後の課題

テスト本願を用いた類似文献検索を行い、各技術分野（テーマコード）毎に再現率の集計を行った結果を図２に示す（図中の X は X 文献を正解文献とした場合、X ＋ Y は、X 文献と Y 文献とを正解文献とした場合、(B) は、IPCC シソーラスを適用しない場合、(T) は、IPCC シソーラスを適用した場合を表す。）。

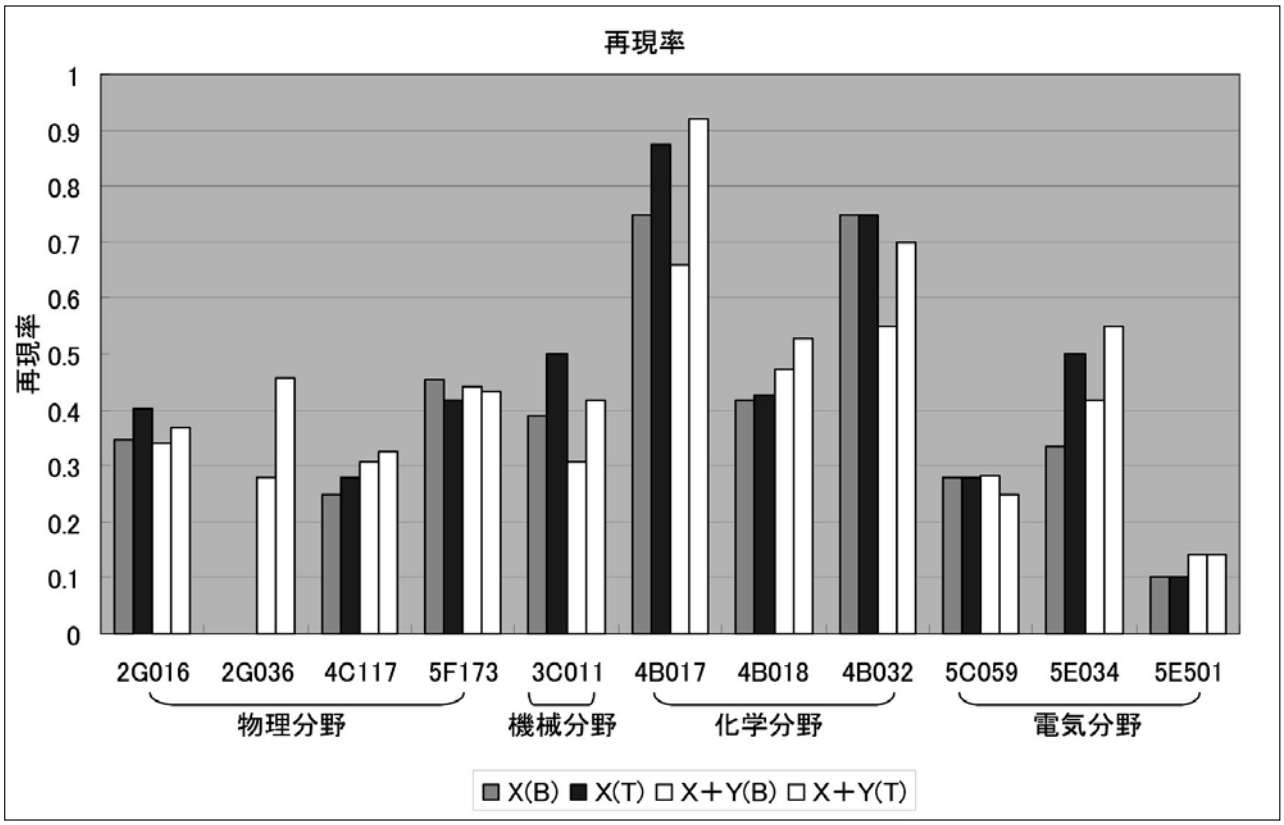

図２　再現率

図２から、各テーマコード毎に、評価したテスト本願の数が少ない（約 10 件）ものの類似文献検索の有効性の傾向が読み取れる（2G036 では、正解文献が X 文献のみの場合の検索結果は、IPCC シソーラスの適用の有無に関わらず上位 100 位外となったため再現率は 0 となった。）。

即ち、検索業務の実務面からみると、再現率は所定数の検索結果の範囲に正解案件が含まれている確率に相当することから、各テーマコードにおける類似検索の有効性が評価でき、類似文献検索が有効な分野の特性を把握することが、類似文献検索システムの今後の活用に繋がる検討課題として挙げられる。

また、IPCC シソーラスの適用の有無を比較すると、定性的には IPCC シソーラスを特徴語の類義語として利用することは、類似文献検索に有効であると考えられる。

但し、汎用的用語辞書 (IPA 辞書) による形態素解析（単語抽出）のままでは、適用した IPCC シソーラスの単語が一般的な単語に分解されてしまう場合がある等の技術的課題があり、今後も継続して調査研究する必要がある。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し、分解された各単語が本来の特徴語の意味と異なることから、意図しない文献（検索時にノイズとなる文献）が上位に多くランクインする可能性があり、今後の検討課題として挙げられる。

参考文献

[1] 高野明彦 , 西岡真吾 , 今一修 , 岩山真 , 丹羽芳樹ほか６名 ,「汎用検索エンジンの開発と大規模文書分析

への応用」,IPA 技術発表会 , 2002

[2] 間瀬久雄 , 岩山真 ,「NTCITR-6 Patent Retrieval Experiments at Hitachi 」,NTCIR-6 ワ ー ク ショップ会議 , pp.403-406, 2007

[3] 岩山真 , 佐藤祐介 ,「引用情報に基づく特許文献の重要度算出方式の検討」, 情報処理学会研究報告 , 2006-FI-83, pp.9-16, 2006