Japio

特許審査業務への適用に向けた概念検索精度の評価

株式会社日立製作所システム開発研究所
間瀬　久雄

PROFILE
平成２年（株）日立製作所入社、システム開発研究所に配属。以来、特許や新聞記事、Ｗｅｂページ等を対象とした、分類自動付与、検索、文章要約、テキストマイニング等の日本語処理の研究に従事。2007 年度から特許版産業日本語委員会委員。
hisao.mase.qw@hitachi.com TEL

1 はじめに

特許庁においては、平成 16 年より業務・システム最適化計画を推進している。「出願人、代理人の利便性向上」「世界最高レベルの迅速かつ的確な審査」「業務の抜本的見直しとシステム経費の削減」の三つを目標として掲げており、第一段階として運営基盤システム、第二段階として新検索システムを構築する予定となっている。

世界最高レベルの迅速かつ的確な審査を実現するためには、検索環境の更なる高度化が不可欠である。検索精度の向上に加え、審査済みの文献に関する情報を蓄積して検索時に活用するなどの最新の機能を備えた新検索システムを実現する必要がある。

そこで、新検索システムにおいて実現すべき機能を明確にすることを目的とした調査研究「審査関連情報を活用した次世代検索システム開発に向けた調査」が平成 20 年度に実施された[2]。本調査研究は特許庁の協力のもとで、審査ナレッジの体系的な活用方法の確立に関する以下の３項目の実現性を調査・検討した。

（１）データマイニング技術を用いた基礎調査
（２）概念検索技術への審査ナレッジの適用
（３）自動分類付与技術への審査ナレッジの適用

本稿では、上記３項目のうち、概念検索技術に関する調査研究結果について述べる。ここで言う概念検索とは、文章を入力としてその文章の内容に類似する文献を出力する検索方式である。本稿では特に、審査業務への適用を踏まえた概念検索の精度評価に焦点を絞って述べる。概念検索の精度評価以外の調査と、他の２項目に関する調査内容については、参考文献 [2] を参照されたい。

2 想定する審査業務

本調査研究では、概念検索技術を審査業務に適用することによって、審査業務の効率を向上させることが可能かを見極めることを目的の一つとした。

本調査研究で想定した特許審査業務フローを図１に示す。

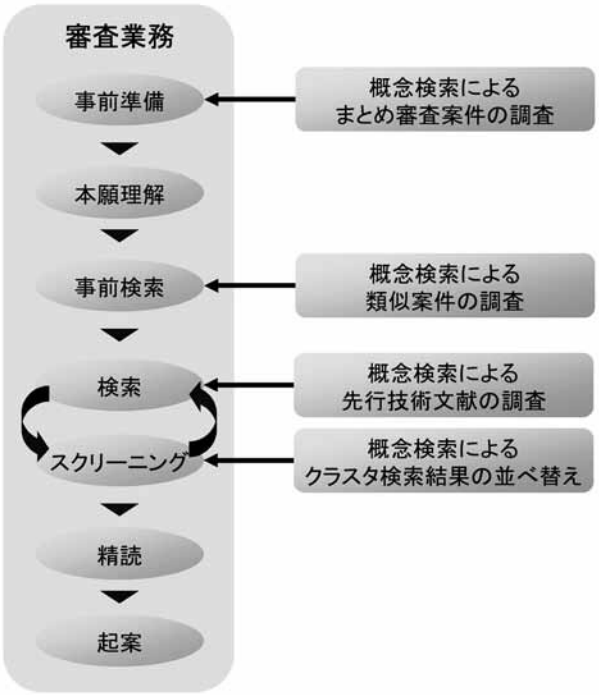

図１　本調査研究で想定する特許審査業務フロー

以下、各業務の概要と、業務における概念検索の果たす役割・意義について述べる。

（1）審査の事前準備（まとめ審査案件の調査）

出願特許を審査する際には、内容の類似する複数の案件をまとめて審査する「まとめ審査」がしばしば行われる。この際、まとめ審査すべき案件を効率良く収集することが課題となる。概念検索によって類似する案件を検索することにより、まとめ審査の案件を効率良く収集できる可能性がある。

（2）事前検索（類似案件の調査）

審査対象となる出願特許（本願）を審査する際には、本願の出願人が過去に出願した類似文献や、本願に類似する過去の文献などを事前に収集する。この収集作業を効率良く行うことが課題となる。概念検索によってこれらの文献を事前に収集することにより、本願理解を促進するとともに、この時点で本願を無効化する先行技術を発見できる可能性がある。

（3）検索（先行技術文献の調査）

先行技術文献を検索する際には、本願や本願の類似文献の内容から検索に用いるキーワードや分類を調査し、サーチ戦略に則って検索を行う必要がある。しかしながら審査官が不慣れな技術分野において、適切な検索キー（FIやFタームなど）を思いつかないような場合には、適当な単語を入力して全文検索を行うことがある。このような場合に、本願中の特定の記載箇所などを入力して概念検索を行うことにより、少数の単語の入力による単なる全文検索に比べて、本願に類似する文献を検索でき、現行のクラスタ検索を補完できる可能性がある。

（4）スクリーニング（クラスタ検索結果並べ替え）

クラスタ検索によって絞り込んだ文献集合を読む作業（スクリーニング）には多大な時間を要しており、作業時間の短縮が課題となる。概念検索によって本願の内容に類似する順に文献集合を並べ替えて表示することにより、所望の文献をより早く発見でき、スクリーニング時間を短縮できる可能性がある。

3 概念検索技術に係る調査研究

3.1 概念検索技術の概要

概念検索とは、文章を入力とし、その文章の内容に類似する文献を検索対象となる文献集合の中から検索し、類似度の高い順に出力する検索方式である。概念検索を用いることにより、複雑な検索式を作成しなくても類似文献を検索できる可能性がある。

概念検索では、図２に示すように、まず入力された文章または文献（本願）から、発明内容を特徴付ける語（特徴語）を自動抽出し、その重要度に応じた重みを自

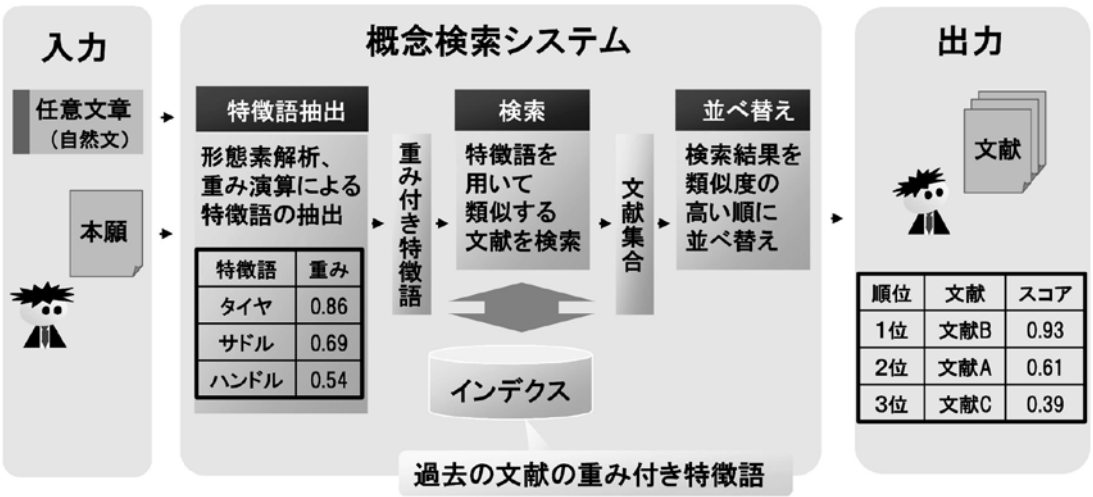

図２　概念検索の概要

動付与する。次に、過去の文献から自動抽出した重み付き特徴語を格納したインデクスと照合することにより、入力文章と各文献との類似度を算出する。最後に、類似度の高い順に文献を並べ替えて、検索結果を出力する。

3.2　概念検索精度向上に向けたアプローチ

概念検索を審査業務に適用するためには、検索精度を向上させることが不可欠である。そのためには、以下の３点の課題をクリアする必要がある。

（１）発明内容を表す特徴語を高精度に抽出する
（２）抽出した特徴語に適切な重みを付与する
（３）特徴語を柔軟に照合し、文献間の類似度を適切に算出する

本調査研究では、このうち（1）および（2）に着目して概念検索精度向上の可能性を調査した。具体的には、以下の３種類の観点に着目し、これらを技術分野毎にチューニングすることによって、精度がどの程度向上するかを検証した（図３）。

（ａ）特徴語を抽出する記載箇所の指定
（ｂ）分類・年範囲の指定
（ｃ）分類の共通性によるスコア補正

3.3　プロトタイプシステム

本調査研究の実施にあたり、概念検索精度検証用のプロトタイプシステムを開発した。検索エンジンとして GETA[3] [1] を、特徴語を抽出するための単語切り出しツールとして ChaSen[4] を採用した。

図４にメイン画面の構成を示す。本画面は以下のエリアから構成されている。

（ａ）書誌情報表示エリア（画面左上）

本願に関する書誌情報として、公開番号、出願人、発明の名称を表示する。「本願表示」ボタンを押下すれば、本願テキストを閲覧できる。

（ｂ）パラメータ設定エリア（画面左中）

概念検索で使用する検索パラメータを設定する。特徴語を抽出する記載箇所、スコア補正方法、特徴語重み付け方法などを設定できる（パラメータ設定の詳細については参考文献 [2] を参照されたい）。

（ｃ）検索条件設定エリア（画面左下）

検索対象文献を絞り込むための検索条件を設定する。年範囲や分類（テーマ、FI、F ターム）によって、検索範囲を絞り込める。また、検索ワードを自然文で入力することにより、本願記載内容を補充できる。

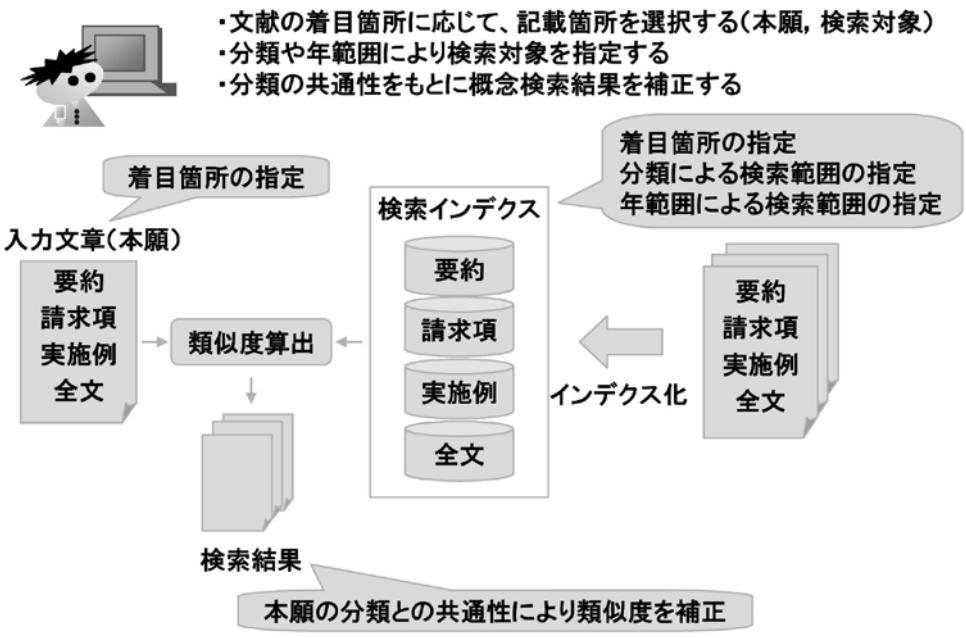

図３　概念検索精度向上に向けた調査観点

1 汎用連想計算エンジン GETA は、独立行政法人情報処理推進機構（IPA）が実施した「独創的情報技術育成事業」の研究成果である

（d）検索結果一覧表示エリア（画面右上）

概念検索結果を表示する。入力文章（本願）と類似する度合いの高い（スコアの高い）文献から順に、公開番号、発明の名称、出願人、概念検索スコアを 10 件ずつ表示する。ユーザは、出力された文献の書誌情報および本文を閲覧して、本願の審査に有用な文献であるか否かを判定する。

（e）引用文献順位表示エリア（画面右下）

本願に対する概念検索精度に関するデータを表示する。正解文献が概念検索結果の何位に出力されたかを表示するとともに、概念検索の精度を評価する指標を算出して表示する。ユーザは、これらの数値データを確認し、検索結果の良し悪しを判定する。

図 4　概念検索プロトタイプシステムのメイン画面

4 概念検索精度評価

4.1　評価内容

本調査研究では、概念検索の精度に関する多くの評価を行っているが、ここでは以下の 2 項目に関する評価結果について報告する。

（1）パラメータチューニングによる精度向上効果

現状の概念検索精度を把握するとともに、技術分野の特性や特許明細書の書き方の特性などを踏まえて、概念検索パラメータをチューニングすることにより、検索精度がどの程度向上するかを評価した。

（2）クラスタ検索結果の並べ替え効果の検証

クラスタ検索によって得られた文献集合を概念検索によって本願に類似する順に並べ替えた時に、正解文献に辿り着くまでに読む文献数がどの程度少なくなるかについて評価した。

4.2　評価環境

1994 年から 2003 年までの 10 年分の公開特許公報および公表特許公報約 370 万件を検索対象として精度評価を行った。対象分野として、以下の 3 技術分野を中心に評価を行った。

（1）機械分野

F16B-F16J(機械要素)、B60K(車両推進装置)

（2）化学分野

A61K(医療用製剤)

（3）電気分野

G11C(半導体メモリ)、G06F(コンピューター一般)

本願全文から抽出した重みの高い特徴語 70 語を使用して概念検索を行った場合の精度を基準（ベースライン）として、検索パラメータチューニング後の検索精度がどの程度向上するかを比較した。

4．3　評価結果（パラメータチューニング効果）

3 技術分野について、特許庁担当審査官が検索パラメータのチューニングを試行錯誤的に行った。具体的には、特徴語を抽出する記載箇所の指定、分類・年範囲の指定、分類の共通性によるスコア補正などを、技術分野の特性に応じて最適になるように設定した。

これらのチューニングによる概念検索精度向上効果を図 5 に示す。チューニングによって、再現率[2] は 29 ～ 56%となり、技術分野毎に精度差が見られるものの、ベースラインの再現率（9 ～ 48%）に比べて、8 ～ 19 ポイント向上した。また、再現率の向上に貢献したパラメータは、技術分野によって異なった。化学分野（医療用製剤）では、実施例に出現する特徴語を重視することにより、再現率の向上が見られた。一方、機械分

2 再現率は、所望の文献を漏れなく検索する割合を示す評価指標である。本稿でいう「再現率 50%」とは、例えば本願の引用文献が 2 件ある場合に、検索結果上位 50 件の中に平均 1 件（1/2=50%）含まれていることを示す。

野（機械要素、車両推進装置）や電気分野（半導体メモリ、コンピューター一般）では、逆に再現率が低下した。また、電気分野では、本願に付与されたＦＩを持つ文献の検索スコアを高く補正することにより、再現率が大幅に向上した。

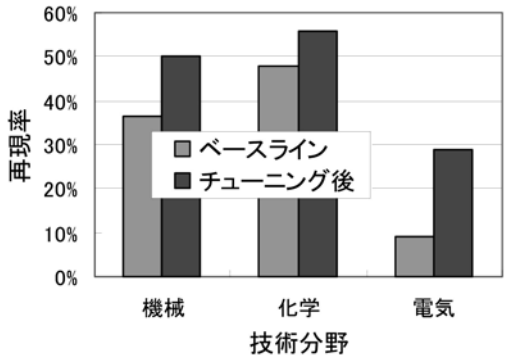

図５　検索パラメータのチューニング効果

上記のチューニング効果が他の技術分野にも当てはまるかを検証した。ランダムに選定した 100 技術分野（テーマ）について、これらの技術分野のいずれかに属する本願（13,812 件）に対する概念検索精度を評価した。

評価結果を図６に示す。このグラフは、100 テーマ全体について、検索出力文献数を増やしていった時の再現率の推移を示している。検索パラメータのチューニングにより、再現率が向上しているのが分かる。また、ベースラインでは、検索結果上位 50 件での再現率が 28% であるのに対して、チューニング後では上位 40 件で同等の再現率を達成でき、同じ精度を出すために必要となる出力文献数を 10 件程度少なくできることが分かった。

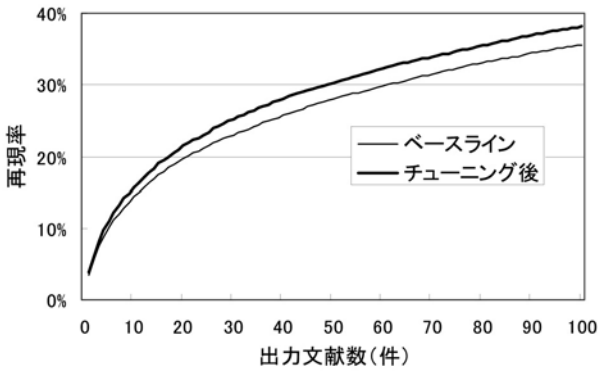

図６　出力文献数と概念検索精度

4.3　評価結果（クラスタ検索結果の並べ替え効果）

３技術分野に属する計 41 件の本願（機械 26 件、化学５件、電気 10 件）に対するクラスタ検索結果文献集合を担当審査官が作成し、この文献集合を概念検索によって本願に類似する順に並べ替えた。クラスタ検索による出力順序（出願日の新しい順）と概念検索による出力順序を比較し、正解文献が何位に出力されるか、またどちらがより上位に出力されるかを比較した。

正解文献の出力順位の平均値を比較した結果を図７に示す。概念検索によって並べ替えることにより、正解文献の出力順位が上昇しており、正解文献にいち早く辿り着けるようになることが分かる。

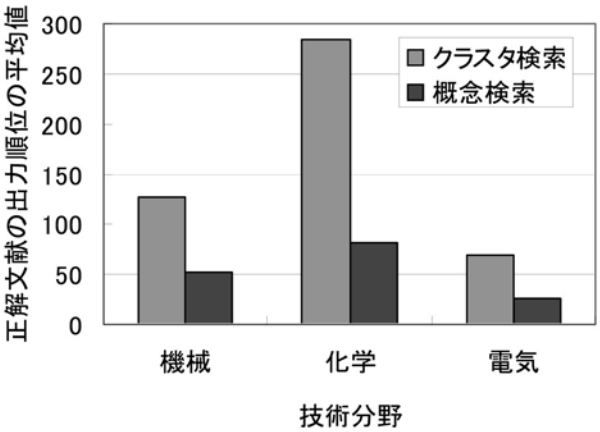

図７　正解文献の平均出力順位の比較

また、概念検索による並べ替えにより、クラスタ検索に比べて出力順位が変動した正解文献の割合を図８に示す。並べ替えによって順位が上昇する正解文献の方が低下する文献よりも多く、６割強から７割強を占めた。

さらに、概念検索による並べ替えによって変動する出力順位の変動幅の平均値を図９に示す。クラスタ検索の出力順位に比べて出力順位が向上する場合の向上幅は非常に大きく、逆に出力順位が低下する場合の低下幅は非常に小さいことが分かる。

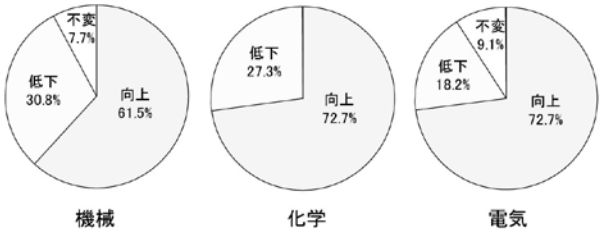

図８　並べ替えで順位が変動した正解文献の割合

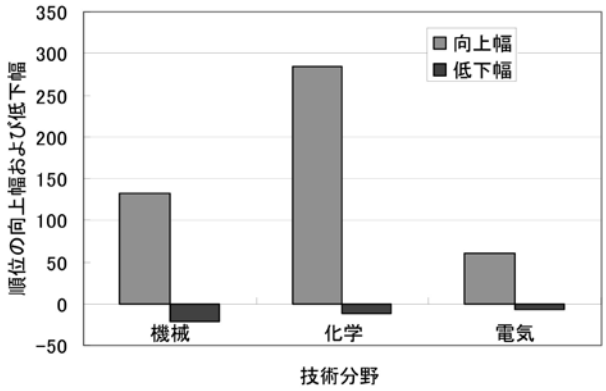

図９　並べ替えによる正解文献の向上幅と低下幅

5 おわりに

本調査研究によって、審査業務に概念検索を適用することが可能であることが分かった。しかし、審査業務においては、概念検索は現行のクラスタ検索に直ちに置き換わるものではなく、適切に使い分けることによって互いに補完していくものであると考える。

概念検索の適用効果をより高くするためには、いくつかの技術課題を解決する必要がある。

検索精度の観点からは、本願の発明内容にドンピシャの文献（X文献）はもちろんのこと、本願に部分的に記載された発明内容に類似する文献（Y文献）を高精度に検索する方式の確立が必要である。そのためには、特許明細書に記載された発明の構成要素および発明ポイント（発明の新規性・進歩性を表す概念）が何であるかを正確に解析し、過去の特許文献の発明ポイントと柔軟に照合可能とする技術が必要となる。

また、使い勝手の観点からは、審査官が概念検索結果の妥当性を評価するための仕掛けが必要となる。すなわち、検索スコア（類似度）の示す意味の明確化や、なぜその文献が概念検索結果として上位に出力されたのか（本願とどこが似ているのか）に関する根拠の提示、さらに、上位何件までを見れば所望の文献が見つかる／見つからないかの見極め支援などである。

今後も、審査業務との親和性のより高い概念検索技術の実現に向けて研究を継続していく。

参考文献

[1] 特許庁業務・システム最適化計画
http://www.jpo.go.jp/cgi/link.cgi?url=/torikumi/system/system_kaitei.htm
[2] 審査関連情報を活用した次世代検索システム開発に向けた調査報告書
http://www.jpo.go.jp/shiryou/toushin/chousa/pdf/kensaku_saitekika/kensaku1_1_2.pdf
[3] 汎用連想計算エンジン GETA
http://geta.ex.nii.ac.jp/geta.html
[4] 形態素解析システム ChaSen（茶筌）
http://chasen-legacy.sourceforge.jp/