類似画像検索と概念検索を統合した特許検索システムの構築

Patent search system using partial image information and text.

株式会社日立製作所　中央研究所　秋良　直人

PROFILE: 2001 年に株式会社日立製作所入社。類似画像検索、自然言語処理等の研究に従事。

株式会社日立製作所　中央研究所　岩山　真

PROFILE: 1992 年に株式会社日立製作所入社。文書検索、自然言語処理等の研究に従事。NTCIR において特許検索用テストコレクションの作成に携わる。2009 年度より特許版産業日本語委員。

TEL 042-323-1111

1 はじめに

一般的な特許検索システムでは、キーワードや特許分類コードの and/or 条件で作成した検索式が検索に用いられる。複雑な検索式を作成することで、検索結果の絞り込みや検索漏れの抑止が可能であるが、検索内容をキーワードで表すことが困難な場合は、粗い絞込みしかできず、大量の検索結果の閲覧に時間を要することが少なくない。

検索式の作成を困難にする理由のひとつが、検索内容を示す自然言語表現に多様性がある場合、キーワードの組合せが無数になることである。このような場合には、キーワードで検索することには限界がある。

キーワードによる検索式の作成が不要な検索方法に概念検索が挙げられる。概念検索は、入力されたテキストと内容が類似している、すなわち、単語分布が類似しているテキストを検索する方法で、特許検索においても有効性が示されている [1]。

概念検索は、複雑な検索式の作成に悩む必要がないものの、用意したテキストに含まれる単語以上の情報がないため、検索内容によっては精度面の限界がある。概念検索のみで上位に期待する結果が得られない場合には、更に別の情報を検索要求として与える必要がある。

特許内容を表わすテキスト以外の情報として、特許図面（以下、図面と呼ぶ）が挙げられる。特許内容を把握するために、図面は貴重な情報である。例えば、発明対象の形状に特徴がある特許であれば、形状を示す図面を確認するだけで内容の把握が可能である。また、類似した発明であれば、類似した観点の図面が含まれていると考えられる。

見た目が類似した画像を検索する類似画像検索技術を活用すれば、これら類似した図面を検索できるため、特許検索に適用することで、概念検索のみを用いる場合と比較し、高精度な結果が得られると考えられる。

本稿では、請求項など検索内容を表わすテキストと、検索内容を表わす図面の両方を検索要求とすることで、高精度な検索ができるという仮定のもと、概念検索と類似画像検索を統合した、メディア統合検索方式を検討し、その予備評価を行った。

2 メディア統合検索

メディア統合検索は、明細書中のテキストを用いた概念検索と、図面を用いた類似画像検索を統合した検索方式（以下、メディア統合検索と呼ぶ）である。検索要求の特許から、検索内容が記載されている部分のテキスト

と、検索内容を表している検索対象に含まれていると考えられる図面を選択し、テキストと図面の両方を検索要求として検索する。

メディア統合検索の類似度 $S_{TI}(d)$ は、次式のように、概念検索で取得した類似度に対して、類似画像検索の類似度を加算することで取得する。

$$S_{TI}(d) = S_T(d) + \sum_{i=1}^{N}\sum_{j=1}^{M} S_I(i,j) \quad (1)$$

ここで、$S_T(d)$ は、検索要求のテキストと、検索対象 d の明細書に含まれるテキストとの概念検索の類似度、$S_I(i,j)$ は、検索要求の i 番目の図面と、検索対象 d の j 番目の図面との類似画像検索の類似度、N は検索要求に含まれる図面の数、M は検索対象 d に含まれる図面の数である。

最終的に、$S_{TI}(d)$ が大きい順に検索対象 d をソートした結果が、メディア統合検索の検索結果である。

図 1 の例では、概念検索のみを用いた場合は、ID:891039 の特許が類似度 0.93 で 1 位、ID:459677 の特許が類似度 0.89 で 2 位なのに対し、メディア統合検索を用いた図 2 の例では、類似画像検索の類似度が概念検索に加算されたことで、ID:459677 の特許が類似度 1.7 で 1 位、ID:891039 の特許は類似度 1.3 で 2 位以下と、順位が逆転する。

3 メディア統合検索の評価と考察

3.1 評価方法

メディア統合検索の効果を検証するため、概念検索の検索精度と、メディア統合検索の検索精度を比較する。

概念検索には、連想検索エンジン MANTA[2] を用い、類似画像検索には、大量の図面から見た目が類似する図面を高速に検索可能な類似画像検索システム EnraEnra[3] を用いる。図面には色情報が含まれていないため、類似画像検索には形状特徴量を用いる。

情報検索に関するタスク型国際ワークショップ NTCIR-5 特許検索タスク [4][5] で使用されたフォーマルラン課題データ 619 件の特許公報で指定された指定部分のテキスト（請求項）と、その特許の図面を検索要求として評価する。

検索対象は、特許公報 10 年分（1993 ～ 2002 年）の明細書（約 340 万件）および図面（約 3,600 万個）で、概念検索の検索対象は明細書中のテキスト全文である。

図面は、検索要求の特許に含まれるすべての図面を用いる場合と、検索に貢献する図面（以下、選択図面と呼ぶ）を 1 個用いる場合の 2 通りで評価する。選択図面は、選択図面の選択に主観が入ることを防止するため、事前に評価を実施し、メディア統合検索に最も貢献する図面

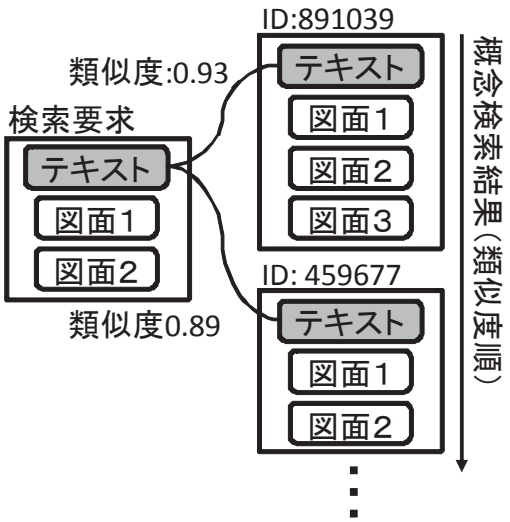

図 1　概念検索の検索結果

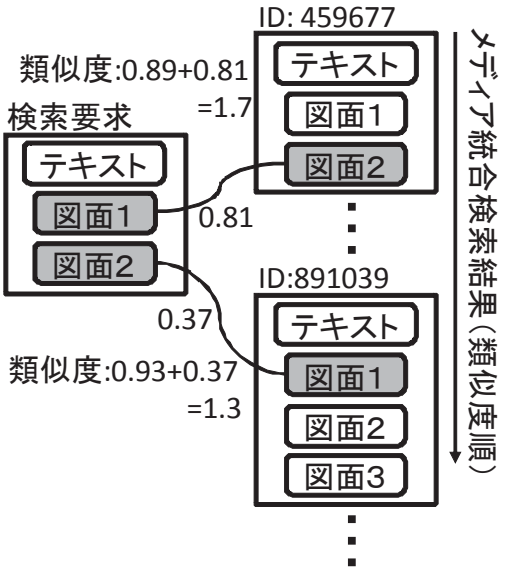

図 2　メディア統合検索の検索結果

を選択図面とする。

3.2　結果と考察

概念検索とメディア統合検索の実験結果を表 1 に示す。検索精度には、特許検索の評価でよく用いられる平均適合率 [6] を用いた。

表 1　メディア統合検索の精度

検索要求	検索方法	平均適合率
請求項（テキスト）	概念検索	0.1115
請求項＋全図面	メディア統合検索	0.0459
請求項＋選択図面	メディア統合検索	0.1186

概念検索を用いた場合の平均適合率 0.1115 に対して、すべての図面を用いた場合のメディア統合検索の平均適合率は 0.0459 と低い結果となったが、選択図面を用いた場合のメディア統合検索の平均適合率は 0.1186 と概念検索と比較し高い結果が得られた。

すべての図面を用いた場合に平均適合率が下がった原因を調査した結果、概念検索の結果で正解よりも下位の特許に、検索要求の図面の酷似図面があり、正解よりも下位の特許の類似度が高くなっていることを確認した。

これは、無条件にすべての図面を用いると、特許内容とは無関係のフローチャートの形状や、大量の特許に共通して含まれているような形状の図面の影響が大きいためであると考えられる。

一方で、選択図面を用いた場合の平均適合率は、概念検索のみの場合と比較し、高い平均適合率が得られた。概念検索と比較し、最も順位変動の差が大きかったのは、飲料缶を包む包装用紙の特許で、包装用紙を展開して 1 枚の紙にした図面が酷似しているために、概念検索では 937 位の順位が、メディア統合検索では、1 位と大きく順位が改善した。

概念検索の順位と比較し、メディア統合検索の順位が大きく改善した上位 20 件の特許では、平均で 292 位の順位が改善し、図面が貢献する特許における有効性を確認した。これらの中には、コピー機の操作パネルの形状、アンテナ形状、農耕機の外観、化学構造式、パチンコ台の形状、会計ソフトの伝票のフォーマット形状、回路図などの類似を確認した。

メディア統合検索の順位が大きく改善した上位 20 位の貢献内容を、図 3 に示す。図面の内容が類似していて順位が改善したのが 55% と最も多く、続いて内容は異なるが化学式の図面が類似しているなど、同じ種別の図面が貢献して順位が改善したのが 20% であった。また、同一出願人の特許が正解の例では、再利用された類似図面が貢献していることを確認した。再利用された図面には、電子データと手描きの違いや、符号の違いなどを確認した。

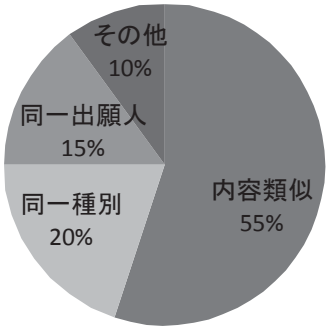

図 3　図面の貢献内容

本稿の評価では、メディア統合検索のベースラインを評価するために、図面の前処理や、図面の種別による重み付けなどを行っていないが、これらの対策を行うことで、更に高い精度が得られると考えられる。

図面の前処理としては、符号など図面の余白にノイズとなる領域を作成してしまう部分の削除などが挙げられる。また、図面の種別による重み付けは、特定分野に偏って頻出するような形状は重みを大きくし、フローチャートのようなどの分野にも含まれる図面の重みを小さくすることなどが考えられる。

4　おわりに

本稿では、テキストと図面を検索要求とするメディア統合検索方式を開発し、その予備評価を行った。NTCIR-5 のデータセットを用いた予備評価で、メディア統合検索の有効性、すなわち図面を特許検索に用いることの有効性を確認した。

特許分野や内容によって、図面の重要性が異なるため、いつでも使える方式ではないが、全文検索、概念検索、メディア統合検索を、同一のシステムで使用できるようになれば、図 4 に示す検索手順の例のように、検索内

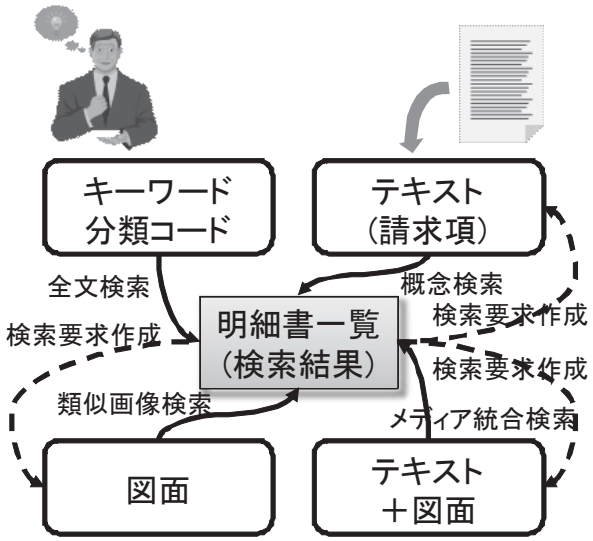

図４　検索手順の例

容に応じて最適な方式を選びながら再検索を実行し、目的の特許に到達することができると考えられる。

本稿では、メディア統合検索のベースラインを評価するために、すべての図面を同一に扱った精度評価を行ったが、図面の種別や内容を反映することで、更に高精度な方式を検討する予定である。

最終的には、特許検索で図面を扱うことが一般的になることを目指したい。

参考文献

[1] 八木，間瀬，岩山．概念検索技術および特許検索への適用可能性について．特技懇，2009.1.30, No.252, 2009.

[2] 安田，今一，岩山，丹羽．連想検索エンジンのスケーラビリティおよび障害耐性の向上．情報処理学会第69回全国大会，2007.

[3] 渡邉，秋良，廣池，松原，平松，永吉，影広，久光．大規模Web画像データベースを用いた画像アノテーションシステムの構築．情報処理学会研究報告，Vol. 2012-CVIM-181, No. 8, 1-6, 2012.3.

[4] N.Kando. Overview of the Fifth NTCIR Workshop. Proceedings of NTCIR Workshop 5 Meeting, 2005.

[5] A.Fujii, M.Iwayama and N.Kando. Overview of Patent Retrieval Task at NTCIR-5. Proceedings of NTCIR Workshop 5 Meeting, 2005.

[6] 間瀬．特許を対象とした概念検索の技術課題．Japio YEAR BOOK 2010, pp. 200-207, 2010.