検索質問の構造解析に基づく多言語類似特許検索システム

藤井 敦　　山田有香　　石川徹也
筑波大学　図書館情報大学　筑波大学
fujii@slis.tsukuba.ac.jp

1 はじめに

近年，知的な創造の成果を活用して産業の国際競争力を強化する動きがある．日本では「知的財産基本法」が2002 年 11 月に成立し，知的財産に関する様々な施策が実施される予定である．

知的財産権の 1 つである特許権は，高度な発明の保護を目的としている．ある着想が発明として具現化し，さらに特許権として成立する過程では，様々な調査が行われる．技術動向の把握に関する調査や対象案件の発明に新規性があるかどうかの調査などがある．

これらの調査では，特許を中心とした大規模な文書群を対象にするため，特許検索システムの利用が有効である．調査の目的によって検索の性質が異なるため，既存の特許検索システムは，特許番号，分類番号，キーワードなどによる多様な検索機能を提供している．

本研究は，請求されている権利を棄却するために行う「無効資料調査（invalidity search）」に焦点を当てる．具体的には，以下のような調査がある．

- 審査請求された発明に対して，特許庁審査官が特許としての適否を審査するために行う実体調査
- 企業知財部のサーチャーが他者の権利を無効化するために行う社内調査

日本は先願主義を採用しているため，対象案件の発明が公知の事実であることを示す資料を検索することが目的である．国内外の特許，論文，Web 文書などが調査対象になる．

無効資料調査では，明細書において権利が請求されている「請求項（クレーム）」などから，審査官やサーチャーが人手でキーワードを抽出し，さらに特許分類などを用いて検索式を構成する．外国語文書も調査する場合は，検索キーワードの翻訳を行う．これらは，対象分野に関する高度な知識が要求される高価な作業である．

本研究は，無効資料調査における種々の作業を自動化することを目的とし，明細書を入力すると，その内容に類似した国内外の文書を横断的に検索する類似文書検索システムを提案する．ただし，検索された文書が実際に無効資料として使えるかどうかの最終判断には法律の知識も必要とされるため，人間が行うことを前提とする．

本研究は，商業的な価値だけでなく，学術的にも意義がある．情報検索の研究では，類似文書検索が中心的に扱われることは稀である．適合フィードバック [3] は，初期検索によって得られた文書を質問として再検索を行うため，間接的に類似文書検索を行う．本研究は，類似文書検索を中心的なテーマとして扱い，その本質を明らかにすることが目的である．さらに，適合判定を用いることなく検索システムを評価する手法を提案する．

2 提案する検索システム

2.1 概要

既存の特許検索システムでは，キーワード，分類番号，日付などの検索キーを AND や OR などの論理演算子で結合して検索式を構成する．近年は，類似文書検索（「概念検索」とも呼ばれる）も実用化されている．この機能を使うと，文書を質問として入力し，その内容に類似する別の文書を検索することができる．論理式による検索は完全一致（exact match）に基づく方式であり，類似文書検索は最良一致（best match）に基づく方式である．

本研究で提案するシステムは最良一致に基づく類似文書検索を行う．本システムの特長は，入力する文書（検索質問）の構造解析を行うことで検索精度を向上させる点にある．入力文書とは，特許の明細書である．また，本システムにおける「入力文書の構造解析」とは，以下に示す 2 つの意味を持つ．

まず，特許請求の範囲である請求項の構造解析がある．1 つの請求項は複数の要素で構成される．構成要素とは，機械の部品，化合物を構成する物質，発明の特徴を表す観点などである．請求項の構造を解析して構成要素に分割することで，発明の本質を明らかにすることができる．

次に，明細書の構造解析がある．1 つの明細書は，請求項以外にも種々の項目を含む．請求項では，権利の範囲を広げるために上位概念を用いた抽象表現が使われる．それに対して,「発明の詳細な説明」では請求内容を具体的に記述している．発明の内容を第三者が理解して再現できるように，明確かつ十分に記載することが特許法で義務付けられているからである．

すなわち，1 つの明細書は，同一内容について抽象的な記述と具体的な記述を含んでいる．明細書の構造解析によって対応する箇所を特定すれば，請求されている権利の内容を具体化することができる．その結果，適切な検索を行うことが可能となる．

以上の解析は，人手で行う無効資料調査においても，人間の知識や経験に基づいて行われている．本研究は，このような専門家の技能をモデル化し，計算機上のシステムとして実装した．

2.2 処理の流れ

図 1 は本システムの概要である．現在は日本語を入力言語とし，外国語として英語を対象にしている．しかし，原理的には特定の言語に依存しない汎用的なシステムである．図 1 に基づいて処理の流れについて説明する．

(1) ユーザは，明細書を入力して無効化の対象となる請求項を 1 つ指定する．

(2) 「構成要素解析」によって対象請求項の構造解析を行い，構成要素に分割する．

(3) 外国語文書を検索するために，構成要素を「翻訳」する．機械翻訳の精度は使用する辞書に依存するため，明細書に記載された特許分類番号を用いて分野辞書を選択する．

(4) 「索引語抽出」によって，構成要素とその翻訳から検索キーワードを抽出する．

(5) 「質問拡張」では，明細書の構造解析を行って請求内容を詳述する箇所を特定し，具体的な検索キーワードを追加する．また，既存の擬似フィードバックを併用する．以上の操作によって，構成要素ごとに検索質問が作成される．

(6) 「文書検索」によって構成要素ごとに類似する文書の候補を取得し，適合スコアで順位付けする．特許分類番号で適宜絞り込みを行う．

(7) 「分析」によって類似文書候補を総合的に評価して再順位付けを行い，最終的な類似文書を決定する．

本システムには，請求項や明細書を構造解析するために，(2)，(5)，(7) の機能がある．これらは，既存の検索システムにはない．構成要素の分割を行わないと，検索結果が特定の要素だけに依存することがある．請求項を構成要素に分割することで，構成要素ごとの重要性を考慮することができる．また，どの構成要素が検索の根拠となっているかをユーザに示すことが可能である．以下，2.3〜2.8 節で，(2)〜(7) について個別に説明する．

2.3 構成要素解析

請求項は，日常言語とは記述形式が異なる一種の制限言語で書かれている．そこで，既存の自然言語解析とは異なる解析手法が必要である．本研究では，以下に示す手掛かりを適宜選択して利用している．

- 構成要素は改行や読点によって明示されることがある．そこで，改行や読点によって請求項を機械的に分割して構成要素を抽出する．
- 請求項の記述形式には，順次列挙形式，構成要素列挙形式，ジェプソン形式などがある．これらの形式を規則化して利用する．現在は，Shinmori ら [4] が提案した特許解析ツールを利用している．

2.4 翻訳

翻訳には，特許用の機械翻訳システム PAT-Transer を用いている．分野辞書を切り替えることで訳質が変わるため，入力となる明細書に付与された国際特許分類（IPC）を用いて使用する辞書を選択する．IPC にはセクション，クラス，サブクラスなどの階層があり，サブクラスまでを使用している．IPC サブクラスと分野辞書との対応は人手で作成した．

構成要素単位で翻訳を行うと，請求項全体の文脈が損なわれる可能性がある．他方において，先に翻訳すると，外国語の請求項記述形式に対応して構成要素解析を行わなければならない．現在は，構成要素を解析した後で翻訳を行っている．

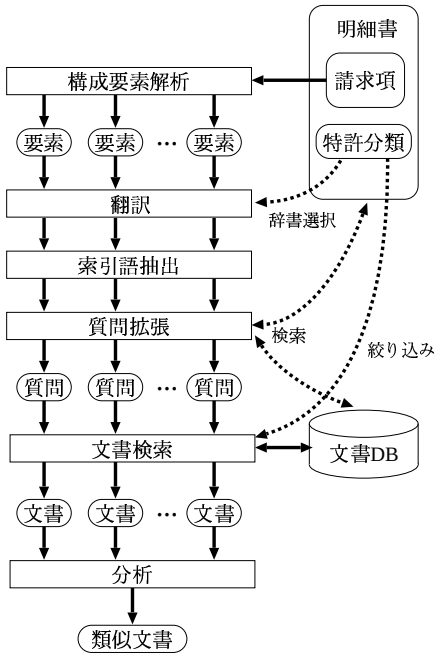

図 1: 多言語類似特許検索システムの概要

2.5 索引語抽出

日本語の構成要素や翻訳された構成要素から，文書検索用の索引語を抽出する．具体的には，形態素解析によって名詞を中心とした内容語を抽出する．また，日本語と英語に対して，特許検索用の不要語リストを事前に人手で作成しておき，リストに含まれる語を削除する．

2.6 質問拡張

質問の拡張は，2 通りの方法で行う．まず，請求項の内容を具体的に記述した箇所を特定して，そこから検索キーワードを抽出する．明細書を墨付括弧（【０００２】など）を基準にして段落に分割し，段落を個別の文書と見なして索引付けを行う．ここでは，図 1 の「文書検索」で使用する検索エンジンを用いる．その結果,「移動体」のような上位概念が「自動車」や「電車」のように具体化される．

丸川ら [6] は，DP マッチングによって請求項と詳細説明の対応付けを行った．それに対して，本システムでは文書検索技術を応用して，請求項に対応する詳細説明を効率良く特定する点が異なる．

もう 1 つの拡張方法は，既存の擬似フィードバックである．1 つ目の手法は入力文書内の情報を用いる手法であり，擬似フィードバックは検索対象文書を用いる手法である．

2.7 文書検索

文書検索のために，既存の手法 [2] を用いて検索エンジンを実装した．原理的には，特許，論文，Web，新聞などの文書ジャンルを問わない．

検索された複数の文書は，適合スコアによって順位付けされる．評価実験に使用した特許公報 5 年分のコレクションは約 170 万文書を含んでおり，ファイルサイズは約 40GB である．このような大規模な文書群に対しても，単体のパソコンを使って，実用的な時間で索引付けや検索が可能である．

検索質問は索引語の列である．また，IPC や日付情報による絞り込みも可能である．無効資料調査では，対象の発明が出願される前に公知であった証拠を探す．そこで，入力された明細書の出願日よりも前に公開された特許公報を検索する．

2.8 分析

構成要素ごとに作成された検索質問を用いて文書検索を行うと，複数の検索結果に重複して含まれる文書が存在する．そこで，検索結果は，図 2 に示すような構成要素と文書を軸とした行列で表現することができる．この図では，8 つの構成要素（1〜8）と 3 つの類似文書候補（A，B，C）が示されている．セル中の数値は文書検索の段階で計算された適合スコアである．

構成要素を横断して文書ごとのスコアを平均することで各文書の最終的なスコアを決定し，再順位付けを行う．A のように多くの構成要素に対する適合スコアが高い文書が最終的な類似文書として優先される．他方において，B は特定の構成要素に対してのみ適合スコアが高いので，最終的な類似文書としての適合性は低い．

ただし，全ての構成要素が等しく重要ではないので，ユーザが不要な構成要素を削除したり，スコアを平均する際の重みを調整するためのインタフェースが有効である．また，構成要素の重要度を自動的に決定する手法が開発されれば，本システムに応用可能である．

3 評価実験

3.1 概要

現在利用できるテストコレクションの制約上，日本語特許で日本語特許を検索する「単言語検索」と，日本語特許で英語特許を検索する「言語横断検索」を異なる方法で評価した．

3.2 単言語検索の評価

単言語検索の評価では，NTCIR-4 特許検索タスクのテストコレクション [5][1] を用いた．当コレクションは，無効資料調査用の検索システムを評価するためのベンチマークであり，以下のデータが含まれる．

- 検索課題: 予備試験用 7 件，本試験用 34 件，追加課題 96 件
- 検索対象: 日本公開特許公報 5 年分（1993〜1997 年）
- 適合判定: 実験当時，予備試験用の適合判定のみが配布されていた．

検索課題は日本公開特許公報から抜粋された公報である．そこで，予備試験用の 7 件を用いれば，単言語検索の評価実験が可能である．ただし，7 件では課題数として少ない．本試験用の適合判定が公開された後に，追加の実験を行う必要がある．

評価尺度として，適合文書の平均順位を使用した．通常，平均精度（MAP）が使用される．MAP は，上位 10 件未満における順位の入れ替わりによって結果が顕著に異なる．それに対して，特許検索では通常数百件の文書を吟味するため，上位 10 件未満における順位の変動よりも，適合文書の順位を 1000 位から 100 位に改善することに意義がある．しかし，MAP ではこのような差異を適切に評価できない．

結果を表 1 に示す．構成要素の解析方法として，解析しない，改行または読点で分割する方法，記述形式を利用する方法 [4] を比較した．質問拡張方法として，拡張しない，明細書を用いた拡張，擬似フィードバック（PRF），明細書による拡張と PRF の併用を比較した．さらに，IPC による絞り込みの効果を評価した．

文書検索手法の性質上，質問拡張をしない場合は，構成要素分割の有無や分割方法に関わらず，結果は同じになる点に注意を要する．

総じて，改行または読点による構成要素解析，拡張方法の併用，IPC による絞り込みが効果的だった．これらを全て適用した場合に平均順位は 203 となり，最も良い結果となった．

3.3 言語横断検索の評価

言語横断検索の評価では，NTCIR-3 と NTCIR-4 の特許検索タスクテストコレクションを併用した．課題には NTCIR-4 本試験用の 34 件を用いた．検索対象文書として，NTCIR-3 の特許和文抄録（JAPIO 抄録）と特許英文抄録（PAJ）を併用した．これらは，1995〜1999 年に公開された特許公報の抄録で，日英対訳コーパスである．ここで問題となるのは，適合判定がないために検索精度を評価できない点である．

そこで，言語横断検索によって英文抄録を検索した結果が「和文抄録を検索した単言語検索の結果にどの程度近いか」という尺度によって評価した．

[1] http://www.slis.tsukuba.ac.jp/~fujii/ntcir4/cfp-ja.html

構成要素		類似文書の候補		
ID	本文	A	B	C
1	映像を処理してパソコン画面上に動画像を表示させるパソコン用動画像処理装置において、	400	600	200
2	映像入力チャンネルからの NTSC 信号を色相別デジタル輝度信号 ... NTSC 信号変換部と、	100	0	100
	...			
8	ことを特徴とするパソコン用動画像処理装置。	300	0	50

図 2: 構成要素と類似文書候補を軸とした行列の例

Carbonell ら [1] は，単言語検索の結果一覧に含まれる上位 N 件を全て適合文書として見なして，言語横断検索の精度を評価した．しかし，N の決め方が困難である点や，上位 N 件の文書を順位によらず全て等価に扱うため評価の厳密性に欠ける点が問題である．

そこで，本研究は単言語検索と言語横断検索の結果を「順位相関」によって比較する評価方法を提案する．順位相関とは，複数のリストにおいて順位の入れ替えがどの程度発生しているかを表す係数である．今回は，ケンドールの τ を用いて順位の入れ替わりが平均して何回発生したかを評価した．ケンドールの τ は [−1,1] の範囲を取り，順位が完全に一致した場合は 1 を取る．

結果を表 2 に示す．言語横断検索では，構成要素解析と明細書による質問拡張の評価をまだ行っていない．そこで，擬似フィードバックと IPC の有効性についてのみ考察する．擬似フィードバックによる質問拡張と IPC による絞り込みを併用した場合に順位相関係数が最高になり，単言語検索の結果に最も近付いた．言語横断検索でもこれらの手法が効果的であった．ただし，複数の手法間での相対的な評価であり，検索精度の絶対値を検証することはできない点に注意を要する．

表 1: 単言語検索の評価結果（適合文書の平均順位）

		構成要素解析		
質問拡張	IPC	なし	改行と読点	記述形式
なし	なし	431	431	431
	あり	385	385	385
明細書	なし	428	367	419
	あり	369	301	360
PRF	なし	361	433	409
	あり	335	358	344
併用	なし	382	267	351
	あり	246	**203**	285

表 2: 言語横断検索の評価結果（順位相関係数）

質問拡張	IPC	順位相関係数
なし	なし	−0.0521
	あり	0.1094
PRF	なし	−0.028
	あり	**0.1255**

4 おわりに

特許調査における無効資料調査の自動化を目指して，多言語対応の類似文書検索システムを実装し，実験によって有効性を評価した．特許以外でも，入力文書が長くなれば複数の話題が混在する可能性が高くなるため，入力文書の構造を解析して検索に反映させることが重要である．提案した検索システムが特許以外の検索に応用可能であるかどうか，今後明らかにする必要がある．

謝辞

新森昭宏氏（インテック・ウェブ・アンド・ゲノム・インフォマティクス）には特許解析ツールを使用させて頂きました．神嶌敏弘氏（産業技術総合研究所）からは貴重なコメントを頂きました．株式会社クロスランゲージには PAT-Transer を使用させて頂きました．

参考文献

[1] Jaime G. Carbonell, Yiming Yang, Robert E. Frederking, Ralf D. Brown, Yibing Geng, and Danny Lee. Translingual information retrieval: A comparative evaluation. In *Proceedings of the 15th International Joint Conference on Artificial Intelligence*, pp. 708–714, 1997.

[2] S.E. Robertson and S. Walker. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 232–241, 1994.

[3] Gerard Salton and Chris Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, Vol. 41, No. 4, pp. 288–297, 1990.

[4] Akihiro Shinmori, Manabu Okumura, Yuzo Marukawa, and Makoto Iwayama. Patent claim processing for readability: Structure analysis and term explanation. In *Proceedings of the ACL-03 Workshop on Patent Corpus Processing*, pp. 56–65, 2003.

[5] 藤井敦, 岩山真, 神門典子. NTCIR-4 における類似特許検索テストコレクションの構築. 情報処理学会研究報告, 2004-NL-159, pp. 45–52, 2004.

[6] 丸川雄三, 岩山真, 奥村学, 新森昭宏. ローカルアラインメントを用いたテキスト間の柔軟な対応付け. 情報処理学会研究報告, 2002-FI-68, pp. 23–28, 2002.