特許，論文間の引用関係を用いた論文用語の特許用語への変換

釜屋英昭[1]，難波英嗣[1]，相沢輝昭[1]，新森昭宏[2]，奥村学[3]
1 広島市立大学 情報科学部
2 インテック・ウェブ・アンド・ゲノム・インフォマティクス
3 東京工業大学 精密工学研究所

1. はじめに

無効資料調査とは，出願された技術が特許権の取得に該当するかどうかの判断をするために，特許庁の審査官が行う審査で，過去に同様の出願技術が存在していたかどうかを調査するものである．無効資料調査を行うには，審査官やサーチャーは，特許と論文データベースの両方を個別に検索する必要がある．しかし，特許では請求範囲をなるべく広く確保するため，一般性の高い特許用語を用いて記述する傾向にある．このため，単純に表層的な単語の一致度を見るだけである従来の検索モデルでは，同じキーワードで特許データベースと論文データベースを検索しても，用語の使われ方の違いから，そのキーワードに関する論文や特許を十分に収集できるとは限らない．そこで本研究では，与えられた論文用語(例えば，DRAM)を特許用語(例えば，半導体記憶装置)に自動変換する手法を提案する．

本研究では，論文用語の特許用語への変換を実現するため，特許と論文間の引用関係に着目する．難波は，ある専門用語を入力すると，それに関連する用語を自動収集する方法を提案している[難波 2005]．この手法では，まず，ある用語を表題に含む論文を収集し，次に，それらと直接引用関係にある論文の表題から用語を抽出し，最後に，それらを頻度順に並べて出力している．本研究でも同様に，ある用語を表題に含んだ論文を収集し，それらと直接引用関係にある特許から，特許のトピックを示す用語を抽出すれば，入力された論文用語に関連する特許用語の変換が実現できると考えられる．そこで，この手法を，特許，論文間の引用関係データベースに適用し，その有効性を実験により検証する．

本論文の構成は以下のとおりである．次節では，関連研究について述べる．3 節では，論文用語の特許用語への変換手法を提案する．4 節では，提案手法の有効性を調べるために行った実験について述べる．

2. 関連研究

これまでに特許を対象とした数多くの検索システムが構築されてきたが[岩山 2001，2003]，近年では特許だけでなく，学術論文も横断的に検索できるシステムの開発やサービスの提供が始まっている．Thomson 社の ISI CrossSearch では，様々な分野の学術雑誌，国際会議の会議録，世界 40 ヶ国の特許発行機関から収集した特許データベースなどを検索することができる．一方，富士ゼロックス社の DocuPat では，日米特許データ 1,800 万件と科学技術振興機構(JST)が提供する科学技術文献データ 2,000 万件を一つのインタフェースで検索することが可能である．しかし，これらのサービスでは特許と論文用語の変換機能は提供されていないため，あるテーマに関する特許と論文を網羅的に収集するには，ユーザ自身が特許と論文用語の違いの問題を解決する必要がある．

この問題に対し，我々はこれまでに，用語の変換とは別の側面から取り組んできた．近年，特許中で関連論文を，逆に論文において関連特許を引用するケースが増えている．このような文書間の引用関係をたどれば，論文や特許と関連する文書を集めることができる．そこで，我々は特許と論文間の引用関係の解析に取り組んできた [安善 2005, 2006]．ただ，現状では，特許中の引用文献の中で論文が占める割合と，論文中の引用文献の中で特許が占める割合は数パーセント程度であるため，あるテーマに関する特許と論文を網羅的に収集するのに，引用関係をたどるだけでは限界があると思われる．そこで，特許，論文間の引用関係に加え，論文用語の特許用語への変換にも取り組み，特許，論文データの効率的な検索環境の構築を目指す．

3. 引用関係を利用した論文用語の特許用語への変換

本節では，まず 3.1 節で，論文用語の特許用語への変換手順について説明する．次に，この変換を実現する上で検討課題となる特許用語の抽出について 3.2 節で述べる．

3.1 論文用語の特許用語への変換手順

本研究では，安善ら[安善 2005, 2006]の手法で得られた特許，論文間の引用データを用い，以下の手順で，論文用語を特許用語に自動変換する．

1. システムに論文用語を入力する．
2. システムは，入力された用語を表題に含む論文をデータベースから検索する．

3. 手順 2 で検索された論文と引用関係にある特許を収集する．
4. 手順 3 で収集された特許から用語を抽出し，頻度順にならべ，出力する．

ここで，手順 4 において，特許中のどの個所から用語を抽出するのかを検討する必要がある．次節では，特許用語の抽出手法について述べる．

3.2 特許用語の抽出

特許から用語を抽出する際, 請求項に着目する．請求項とは，「特許を受けようとする発明を特定するために，必要と認める事項のすべてを記載した項」のことであり，特許明細書の中で最も重要な個所である．また，この個所は，請求範囲をなるべく広く確保するため，一般性の高い特許用語を用いて記述されるという特徴がある．そこで，本研究では，請求項から用語を抽出する．

さて，図 1 は，請求項の一例であるが，この例から分かるように，請求項は慣例的に長い１文で記載されるため，請求項すべてから用語の抽出を行うと，その中に不要な語が多く含まれてしまう．

> 操作手段によりアクチュエータを駆動して所望の作業を行う**作業機**において, 前記作業の作業機構に作成する負荷を検出する負荷検出手段と，この負荷検出手段の検出値に応じた周波数の信号を出力する第１の周波数変換器と，当該負荷検出手段の検出値に応じた周波数のパルスを出力する第２の周波数変換器と，前記第１の周波数変換器から出力される信号を前記第２の周波数変換器からのパルスの出力期間だけ間欠的に出力する変調手段と，この変調手段の出力に応じて振動を発生する振動発生手段とを設けたことを特徴とする**作業機の操作用仮想振動生成装置**

図 1 請求項の例(特開平 10-011111 より引用)

ここで，請求項には以下に述べるような 2 つの構造的な特徴が存在する[新森 2004]．一つ目は，請求項の記述末尾に名詞または記号が存在し，その直前に名詞があり，さらにその直前に名詞，記号，または助詞「の」が連続的に出現して「名詞のまとまり」を形成する，という特徴である[1]．二つ目は，「において,」や「であって,」などの文字列を用いて記述を前半部と後半部に分割するとき,「において,」や「であって,」の直前にも，記述末尾と同様の「名詞のまとまり」が存在する，という特徴である[2]．このまとまりは，発明の名称を表していることが多い．新森らは，手がかり語を用いて請求項の構造を解析する手法を提案しているが，この解析結果を用い，「名詞のまとまり」から用語の抽出を行う．

本研究では，この他，特許中の請求項間の関係にも着目する．特許中には，複数の独立請求項(他の請求項を引用しない請求項)と, 各独立請求項を引用する従属請求項が存在する．また，一般的に独立請求項では上位概念で，従属請求項では下位概念で発明が記載される．このことから，用語抽出の対象となる請求項を，独立請求項とそれを引用する従属請求項に限定した方が，特許中のすべての請求項を使うより良い抽出が可能であると考えられる．一方，一般性の高い特許用語を抽出するには，独立請求項のみを抽出対象にした方が良いと考えることもできる．そこで，独立請求項を使った場合，独立請求項とその従属請求項を使った場合，特許中のすべての請求項を使った場合のそれぞれで実験し，結果を比較する．なお，今回は, 独立請求項として第一請求項(特許中にある複数の請求項の中で，一番最初に記載されているもの)を用いる．

4. 実験

3 節で述べた手法の有効性を調べるために実験を行った．

4.1 実験手法

比較手法

以下の 6 通りで特許用語を抽出し，結果を比較する．以下，(1)～(5)は提案手法で，いずれも特許，論文間の引用関係を利用した抽出方法である．また，これらの手法と比較するため，入力された用語と共起頻度の高い用語を出力する方法をベースラインとした．なお，ベースラインシステムでは，汎用連想検索エンジンGETA[3]を利用する．

(1) 第一請求項から名詞句を抽出
(2) 第一請求項を構造解析し，名詞句を抽出
(3) 全請求項から名詞句を抽出
(4) 全請求項を構造解析し，名詞句を抽出
(5) 第一請求項とその従属請求項を構造解析し，名詞句を抽出
(6) 与えられた論文用語と高頻度で共起する名詞句を抽出（ベースライン）

[1] 図 1 の場合「作業機の操作用仮想振動生成装置」
[2] 図 1 の場合「作業機」
[3] http://geta.ex.nii.ac.jp

実験に用いるデータ

実験には特許公開公報(1993～2002 年)を用いる．特許，論文間の引用関係データは，安善の手法[3]を用いて抽出した特許中の引用論文の書誌情報約 85,000 件を用いる．

正解データセット

正解データセットは以下の手順で作成した．

1. 特許中で引用されている論文の書誌情報 85,000 件中から名詞句を抽出し，頻度順に並べる．
2. その中から論文用語 25 語を人手で選択する．
3. 論文用語毎に「比較手法」で述べた手法 3 を用いて請求項中のすべての名詞句を抽出し，頻度順に出力する．
4. その中から人手で正解判定を行う．

手順 2 で選択された論文用語の一部を以下に示す．

CPU，半導体レーザ，DRAM，メモリセル，ワードプロセッサ，ノボラック樹脂，CD

なお，正解判定を行う際，以下の点を考慮した．

[基準 1] 概念的に最も近い用語のみ正解
例えば，「ワードプロセッサ」という論文用語に対して，「文書編集装置」を正解とし，ワードプロセッサの構成要素である「表示装置」は不正解とした．

[基準 2] 特許データベース中の文書頻度
ある用語の文書頻度が特許データベース中で極端に低い場合は，その用語は特許検索を行う上で有用でないと考え，不正解とした．

[基準 3] 基準 1 で選択されたものとの比較
ある用語が基準 2 を満たさない場合でも，その用語が基準 1 で選択されたものと概念的にほぼ等しいと判断される場合，低頻度でも正解とした．例えば，「ワードプロセッサ」に対して，「文書編集装置」と概念的にほぼ等しい「文書作成装置」も正解である．「レーザ」と「レーザー」のような表記のゆれについても，一方が正解と判定されていれば，もう一方も正解とした．

評価尺度

評価には，以下に定義される ε という尺度を用いる． これは，質問応答システムの評価において一般的に用いられる MRR(mean reciprocal rank)を拡張したものである[清田 2004].

$$\varepsilon = \frac{\sum_{i \in R} \frac{1}{i}}{\sum_{j \in \{1,2,\cdots n\}} \frac{1}{j}}$$

ここで，n は入力に対する正解の数，R は出力されたリスト中の正解順位番号の集合である．ε は正解がすべて最上位に順位付けされたときに，最大値 1 をとる．

不要語句の削除

「方法」や「記載」といった用語は，分野を問わず多くの特許請求項中に出現する．このような用語を出力する特許用語から除外するため，不要語句リストを作成した．このリストの作成は，特許 10 年分に含まれる名詞句を文書頻度順に並べ，頻度の高いものの中から不要と思われる語句を人手で選択することで作成した．以下に不要語句の例を示す．

方法，記載，発行，文献，使用，利用，
詳細，製造，提案，製造方法，データ
(計 350 個)

4.2 実験結果

実験結果を表 1 に示す．

表 1：実験結果

提案手法					ベースライン
(1)	(2)	(3)	(4)	(5)	(6)
0.108	0.171	0.149	0.224	0.231	0.011

表 1 から分かるとおり，提案手法はすべてベースラインを上回った．また，提案手法の中では，手法 5 が最も優れていた．

4.3 考察

まず，請求項の構造解析の有効性について考察する．手法 1 と 2 を比較すると，構造解析を用いた手法 2 の方が優れていることがわかる．また，手法 3 と 4 を比較すると，やはり構造解析を用いた手法 4 の方が勝っている．このことから，請求項の構造解析が特許用語の抽出に有効であることがわかる．

次に，請求項間の関係を考慮することが有効であるかどうか，検討する．手法 2，4，5 の結果を比較すると，第一請求項とその従属請求項を用いた手法 5 が最もすぐれており，第一請求項しか用いない手法 2 が最も悪い結果となった．手法 2 の結果が悪い原因は，抽出個所の制限が強すぎ，ノイズを減らすだけでなく，抽出できた正解の数も減ってしまったためである．全請求項を使った手法 4 は，手法 2 と比べると抽出できた正解の数は

大きいものの，不正解のものも数多く抽出してしまっているため, 手法 5 に劣る結果となっている．

各手法において，１つの特許から名詞句を抜き出す際に，請求項をいくつ用いているのか調べたところ，表 2 のような結果になった．手法 5 は，手法 4 と比べ，抽出対象となる請求項の数が 4 割未満であるにもかかわらず抽出精度が手法 4 よりも高くなっていることから，第一請求項とその従属請求項内に，高い確率で正解が含まれていることがわかる．

表 2：各手法における請求項の数の平均

手法	第一 請求項	すべての 請求項	第一請求項と その従属請求項
請求項 数	1	6.75	2.53

今回最も精度のよかった手法 5 についてエラー分析を行った．その結果，エラーには大きく次の 2 つの理由があることがわかった．

- **入力語句と関連性が低い，あるいは一般的な用語の抽出**

例えば「CD」という入力に対し，「屈折」や「再生」といった用語が該当する．「屈折」や「再生」という用語そのものは「CD」とは無関係ではないが，これらは一般的な単語であり，特許検索には不向きである．全体の 70%以上がこの種類の抽出誤りであった．これらは，逆文書頻度(IDF)等を考慮することで改善できると考えられる．

- **正解用語と部分文字列が一致している用語の抽出**

例えば，「DRAM」という入力に対して，「半導体記憶装置」が正解であるが，この正解と部分文字列が一致する「半導体基板」や「半導体装置」が抽出されていた．この誤りは抽出誤り全体の約 15%を占めていた．この誤りの場合は，正解語句の上位・下位概念である場合も多く，IDF を適用してもあまり効果的ではないと思われる．現在，具体的な解決策は見つかっていないが，今後検討していく必要がある．

5 おわりに

本研究では, 特許, 論文間の引用関係に着目し，論文用語を特許用語に自動的に変換するシステムの構築を行った．提案手法では，論文用語が与えられると，その用語を含んだ論文を引用する特許を収集し，そこから特許用語を抽出して，頻度順にならべて出力する．その際，特許請求項の構造と請求項間の関係を考慮した．提案手法の有効性を確認するため，実験を行った．その結果，請求項の構造解析が特許用語の抽出に有効であり，また，第一請求項とその従属請求項から特許用語を抽出した場合に最も良い結果が得られた．

6 今後の課題

今回の実験結果から，提案手法のある程度の有効性は確認できたが，実験に用いたデータセットは小規模であるため，この結果のみから提案手法が有効であると断定するまでには至らないと考えている．今後はより大規模なデータセットを用いて評価を行う必要がある．それと，考察で述べた逆文書頻度(IDF)の他にも，提案手法を改良する方法についても検討していく必要がある．

謝辞

本研究について議論していただいた IRD 国際特許事務所の谷川英和氏，デュオシステムズの宮原俊一氏，ピコラボの鈴木泰山氏，日立製作所の岩山真氏に感謝致します．今回実験に用いた特許データは，国立情報学研究所の許可を得て，NTCIR テストコレクションを利用させていただきました．本研究は，NEDO 平成 16 年度産業技術研究助成事業の支援を受けて行われました．

参考文献

[安善 2005] 安善奈津美，難波英嗣，相沢輝昭，奥村学“特許、論文データベースを統合した検索環境の構築” 情報処理学会 研究報告，NL-168，pp.21-26，2005.

[安善 2006] 安善奈津美，難波英嗣，相沢輝昭，奥村学“特許、論文データベースを統合した検索環境の構築” 言語処理学会 第 12 回年次大会，2006.

[岩山 2001] 岩山真, 藤井敦, 高野明彦, 神門典子, “特許コーパスを用いた検索タスクの提案”, 情報処理学会研究報告 2001-FI-63，pp.49-56, 2001.

[岩山 2003] 岩山真, 藤井敦，神門典子，丸川雄三, “特許検索の諸相 ―「NII テストコレクション 3 特許」を用いて―” 言語処理学会第 9 回年次大会, pp.671-674, 2003.

[清田 2004] 清田陽司，黒橋禎夫，木戸冬子“自動抽出した換喩表現を用いた係り受け関係のずれの解消”自然言語処理，Vol.11，No.4，pp.127-145，2004.

[新森 2004] 新森昭宏，奥村学，丸川雄三，岩山真“手がかり句を用いた特許請求項の構造解析”情報処理学会論文誌，Vol.45，No.3，pp.891-905，2004.

[難波 2005] 難波英嗣“論文間の引用情報を利用した関連用語の自動収集”言語処理学会 第 11 回年次大会，2005.