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]　奥村 学[2]　新森 昭宏[3]　谷川英和[4]　鈴木泰山[5]

1 広島市立大学 〒731-3194 広島市安佐南区大塚東 3-4-1

2 東京工業大学　3 インテック・ウェブ・アンド・ゲノム・インフォマティクス

4 IRD 国際特許事務所　5 ピコラボ

E-mail: kamaya@nlp.its.hiroshima-cu.ac.jp

あらまし 近年，特許出願が研究活動のひとつとして重視されるようになってきており，研究者が特許と論文を検索する機会が増えつつある．しかしながら，特許と論文データベースを検索する際に，単純に表層的な単語の一致度を見る従来の検索モデルでは，同じキーワードでそれぞれの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，特許，論文間の引用関係に着目し，論文用語から特許用語へ自動変換(例えば「フロッピーディスク」を「磁気記録装置」に変換)する手法を提案する．実験の結果，提案手法の有効性が確認された．

キーワード 特許，論文，引用関係，上位下位関係

Paraphrasing Scholarly Terms into Patent Terms Using Citation Relations between Research Papers and Patents

Hideaki KAMAYA[1]　Hidetsugu NANBA[1]　Manabu OKUMURA[2]

Akihiro SHINMORI[3]　Hidekazu TANIGAWA[4]　Taizan SUZUKI[5]

1 Hiroshima City University　3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194 Japan

2 Tokyo Institute of Technology　3 INTEC Web and Genome Informatics Corporation

4 IRD Patent Office　5 Picolab

E-mail: kamaya@nlp.its.hiroshima-cu.ac.jp

Abstract Recently, the number of occasions when academic researchers retrieve patents and research papers has been increasing, because applying for patents is considered to be an important research activity. However, retrieving patents using keywords is a laborious task for them because terms used in patents are more abstract than those of research papers in order to extend the scope of claims. Therefore, we propose a method to paraphrase scholarly terms into patent terms (e.g. paraphrase “floppy disc” into “magnetic recording medium”) using citation relations between research papers and patents. We conducted some experiments, and confirmed the effectiveness of our method.

Keyword patent, research paper, citation relations, hypernym-hyponym relations

1. はじめに

近年，知的所有権に対する関心が高まり，企業はもちろん，個人が特許を取得するケースも増加してきている．特許出願の際には過去に同様の出願技術が存在していたかどうかの確認作業が必要不可欠である．特許庁の審査官や企業のサーチャーが行なうこの作業を無効資料調査と呼ぶ．無効資料調査では，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がある．例えば「DRAM」は「半導体記

憶装置」と記述される。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を特許用語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

これまでにも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った．

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が，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.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. 引用関係を用いた論文用語の特許用語への変換

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．

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，強調および下線筆者)

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].

第一には，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」（図 1「作業機の操作用仮想振動生成装置」）を形成する，という特徴である．

第二は，「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」（図 1「作業機」）が存在する，という特徴である．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．我々のこれまでの研究において，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，結果を比較したところ，独立請求項とその従属請求項を使った時に最も高い精度が得られた[釜屋 2006]．そこで，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，最初に記載されているもの)とその従属請求項を用いる．

3.3 Mase 手法を用いた提案手法の改良

特許明細書の「符号の説明」という項目には，「磁気記憶装置(フロッピーディスク)」といった記述が数多く存在する．Mase ら[Mase 2005]は,このような記述から,「磁気記憶装置」と「フロッピーディスク」といった関連用語対を抽出し，特許検索の際の query expansion に利用している．この手法は,「フロッピーディスク」という用語を「磁気記憶装置」という特許用語に変換する本研究においても有効であると考えられる．そこで，Mase 手法を実装し，実際に論文用語を入力して調べた結果，いくつかの入力用語に対しては，3.2 節で提案した手法よりも高い精度で変換できることが確認されたが，入力された用語に対する特許用語が全く見つからないといった場合も多数あった．

ここでは Mase 手法を用いながら提案手法の改良を行なうこととする.ある入力用語に対し，Mase 手法によって，例えば「磁気記憶装置」や「リムーバブル記憶装置」といった出力が得られた場合，入力用語は何らかの装置に関する用語ではないかと考えられる．このような場合，提案手法で得られた結果の中で用語の最後が「装置」で終わっているものは，他の用語よりもスコアを上げて用語の出力順序を変えることで，提案手法の改良を行う．

3.4 上位下位関係を考慮した提案手法の改良

3.2 節でも述べたように，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される．つまり，特許用語の多くは論文用語の上位用語であると考えられる．そこで引用関係を用いた提案手法とは別に，特許シソーラスを用いた上位語の収集による手法を提案する．このシソーラスは「A や B などの C」などの定型表現に着目して，用語の上位,下位概念を自動的に構築したものである．

今回構築したシソーラスでは,「などの」「等の」の 2 種類の定型表現に着目し，特許公開公報(1993～2002 年)から，これらの表現を含む文を収集している．収集してきた文から上位・下位関係は出現頻度で重みをつけ，約 700 万件得ることができた[難波 2007].

このシソーラスを元に，入力語句の上位語にあたる用語を収集し，提案手法の回答候補として追加することで，より網羅的に入力語に対する特許用語が収集できるように提案手法の改良を行う．

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

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 60 語を人手で選択する．
3. 論文用語毎に請求項中のすべての名詞句を抽出し，頻度順に出力したものと，3.4 節で述べた上位語として頻度順に出力されたものを合わせる．
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，光ディスク

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)

4.1.1 論文用語の特許用語への変換実験

比較手法

以下の 4 通りで特許用語を抽出し，結果を比較する．以下，(1)～(3)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，(4)は，ベースライン手法であり，汎用連想検索エンジン GETA[1]を利用して，入力された用語と共起頻度の高い用語を出力する．

(1) 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
(2) (1)に 3.3 節で述べた Mase 手法を用いて改良し，名詞句を抽出
(3) 3.4 節で述べた、与えられた論文用語の上位語と考えられる名詞句を抽出
(4) 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）

4.1.2 提案手法と上位手法の統合実験

提案手法 2 と 3 のそれぞれ出力に対し，上位 1 件のスコアが 1 となるよう正規化し，手法 2 と 3 のスコアを，提案手法 2*λ+提案手法 3*(1-λ)で加える．λの値は 0～1 の範囲で 0.1 ずつ変えて実験を行う．

4.2 実験結果

4.2.1 論文用語の特許用語への変換

実験結果を表 1 に示す．表 1 より，今回提案した 3 つの手法は，ベースラインよりも良い結果が得られていることが分かる．

[1] http://geta.ex.nii.ac.jp

表 1：提案手法とベースライン手法との比較

提案手法			ベースライン
（1）	（2）	（3）	（4）
0.14	0.15	0.20	0.01

4.2.2 提案手法 2 と 3 の統合実験

実験結果を図 2 に示す．図 2 より，λが 0.1，0.3，0.7 の時に，精度が 0.21 と，最も高い値を示していることが分かる．

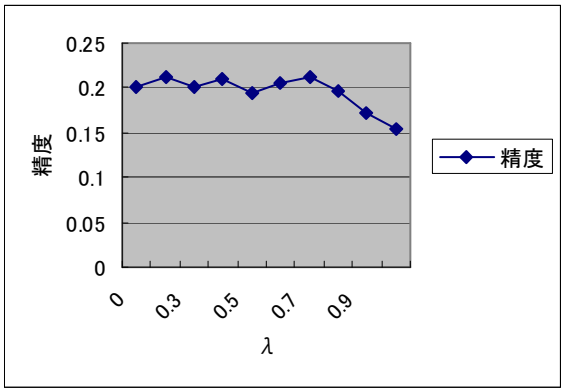

図 2 提案手法 2 と 3 の統合実験の結果

4.3 考察

4.3.1 論文用語の特許用語への変換

第一請求項とその従属請求項を用いた手法 1 と 2 を比較すると，手法 2 のほうが精度が高いため，Mase 手法による改良が有効であることが分かる．

提案手法の中では，引用関係を用いていないシソーラスを用いた結果が一番良い．この理由として，シソーラスを用いた場合，非常によく似た同義語を大量に取ってくるために，正解判定の[基準 3]に基づきすべて正解として追加したためだと考えられる．例えば，光ディスクの正解判定として，「光学的記憶媒体」や「光学式記憶媒体」のような語を正解として追加した．

また，手法 2 と 3 の実際の出力を見てみたところ，それぞれの手法によってのみ正解を抽出できないことがあった．そのような例として，「機械翻訳」と入力した場合，従来の手法では「機械翻訳装置」や「言語変換装置」などが出力されたが，提案手法 3 では「自然言語処理」

や「機械的処理」しか出力されなかった.

4.3.2 提案手法と上位手法の統合実験

表 2 より，λの値が 0~0.8 までは精度にそれほど変化が見られない．しかし 4.3.1 節で述べたようなそれぞれの手法でしか正解を抽出できない問題を，提案手法 2 と 3 の結果を統合することで，精度を保ったまま，より網羅的に入力語に対する特許用語が収集可能になったと考えられる.

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, Mase 手法による改良, 引用関係を用いた手法(提案手法 2)とシソーラス手法(提案手法 3)の統合が有効であるということが分かった.

6. 今後の課題

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．なお，今回は提案手法に上位手法の出力を取り入れる際に，加算する比重を考え合成したが，今後は他の合成方法についても検討していく必要がある.

謝辞

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いた．本研究は，NEDO 産業技術研究助成事業の支援を受けて行われた.

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 ー「NII テストコレクション 3 特許」を用いてー”言語処理学会第 9 回年次大会, pp.671-674, 2003.

[釜屋 2006] 釜屋英昭, 難波英嗣, 相沢輝昭, 奥村学“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”言語処理学会 第 12 回年次大会, pp.723-726, 2006.

[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.

[Mase 2005] Mase H., Matsubayashi T., Ogawa Y., Yayoi T., Sato Y. and Iwayama M. “NTCIR-5 Patent Retrieval Experiments at Hitachi,” Proc. of NTCIR-5 Workshop Meeting, pp.318-323, 2005.

[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.

[難波 2007] 難波英嗣, 奥村学, 新森昭宏, 谷川英和, 鈴木泰山“特許データベースからのシソーラスの自動構築”言語処理学会 第 13 回年次大会，2007.