特許，論文データベースを統合した検索環境および動向分析ツールの構築

○難波英嗣[1)]，釜屋英昭[1)]，奥村学[2)]，谷川英和[3)]，新森昭宏[4)]，
鈴木泰山[5)]，宮原俊一[6)]
広島市立大学[1)]，東京工業大学[2)]，IRD国際特許事務所[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)]，ピコラボ[5)]，デュオシステムズ[6)]
〒731-3194 広島市安佐南区大塚東 3-4-1 広島市立大学 情報科学部
Tel: 082-830-1584 FAX: 082-830-1584
E-mail: nanba@its.hiroshima-cu.ac.jp

Construction of a cross-genre retrieval environment and a technical mining tool by integrating a patent and a research paper database

○NANBA Hidetsugu[1)], KAMAYA Hideaki[1)], OKUMURA Manabu[2)],
TANIGAWA Hidekazu[3)], SHINMORI Akihiro4), SUZUKI Taizan[5)],
MIYAHARA Shun'ichi[6)]
Hiroshima City University[1)], Tokyo Institute of Technology[2)], IRD Patent Office[3)],
INTEC Web and Genome Informatics[4)], Picolab[5)], DUO Systems[6)]
3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194 Japan
Phone: +81-82-830-1584 Fax: +81-82-830-1584
E-mail: nanba@its.hiroshima-cu.ac.jp

【発表概要】

本稿では，特許と論文データベースを統合し，ユーザが特許と論文を横断的に検索したり，技術動向を分析したりすることを可能にするツールについて述べる．本研究では，まず，特許と論文間の引用関係を解析し，次に，この関係を用いて，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開発する．ユーザは，引用関係をたどったり，用語の変換技術を用いたりすることにより，特許と論文を横断的に検索できる．次に，統合されたデータベースから特定分野の文献を収集し，それらからその分野でどのような要素技術がいつ頃から使われたかという情報を自動抽出し，年代順に並べてグラフとして出力する．この結果，ユーザはある分野の技術動向を分析できる．

【キーワード】

特許，論文，引用関係，技術動向分析

1. はじめに

近年，大学研究者にとって，特許出願が重要な研究活動のひとつとして考えられるようになった結果，研究者自身が関連論文だけでなく特許も調べたり，特許を出願したりするという機会が増えている．2006 年 6 月に政府の知的財産戦略本部が発表した「知的財産権推進計画 2006」[1]においても，大学研究における特許情報の重要性が謳われており，大学研究者の利用を想定した特許･論文情報統合検索システムの整備もこの計

[1] http://www.ipr.go.jp/

画のひとつに挙げられていることから，この傾向は今後さらに強まっていくと思われる．

しかし，特許にあまり馴染みのない研究者が特許検索を行うのは容易なことではない．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用い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．また，特許を効率的に検索するには，IPC コード，FI コード，F タームなどの分類記号を使いこなす必要があるが，それには専門的な技術と経験が必要となる．

そこで，本研究では，特許と論文データベースを統合し，研究者にとって特許検索を容易にする環境の構築を目指す．また，この統合されたデータベースを用いて，技術動向の分析を行うツールの開発を行う．

2. 特許と論文データベースの統合

我々は，Web上のPostscriptとPDF形式の日英論文約 78,000 件を収集し，引用論文データベースPRESRI [2]を構築している[1]．さらに，PRESRIと特許間の引用関係を解析し，両者の統合を行っている[2]．特許と論文間の引用関係を用いることで，研究者はキーワード検索で論文を検索した後，引用関係をたどって関連特許を収集できる．

しかし，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にしか過ぎず，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．

[2] http://www.presri.com

そこで，特許と論文間の引用関係に加え，論文用語の特許用語への自動変換にも取り組み(例えば「DRAM」を「半導体記憶装置」に変換する)，特許，論文データの効率的な検索環境の構築を目指している．次節では，この自動変換技術について述べる．

2.1 論文用語の特許用語への変換手順

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．我々は過去の研究において，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[3]．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．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．

以下に，その手順を示す．

1. システムに論文用語を入力．
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．
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．
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．

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．

2.2 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．

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[4]．ひとつ目は，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」（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．ふたつ目は，「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」（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．実際に，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場合において一番良い結果が得られた．

2.3 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[5]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，2.2 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あった．そこで，Mase 手法を用いて提案手法を改良する．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あると考えられる．そこで，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．

2.4 出力例

以下に，「ワードプロセッサ」と「DRAM」を入力した場合の出力結果(スコア順)をそれぞれ示す．また，人間が正解と判定したものを下線で示してある．

入力用語(ワードプロセッサ)

1. 文書編集装置
2. 文書作成装置
3. 文書処理装置
4. 文書作成支援装置
5. 共起関係計算装置

入力用語(DRAM)

1. 半導体記憶装置
2. ダイナミックランダムアクセスメモリ
3. 半導体メモリ
4. 集積メモリ
5. キャッシュメモリ

この結果から分かるように，上位に多くの正解が出力されており，提案手法の有効性が確認できる．

3. 技術動向の分析

ある分野において，「どのような要素技術がいつ頃から使われているのか」という情報を網羅的に収集し整理することは，その分野の技術動向を概観するのに必要不可欠であるが，その作業には多くの時間と労力を要する．そこで，このような情報を自動的に抽出し，可視化するシステムの構築を目指す．なお，今回は論文データのみを対象にし，次節ではその手法を述べる．

3.1 技術動向の抽出手法

技術動向情報を抽出し，可視化するには，まず，特定分野の論文を収集し，次に，そこから可視化に必要な情報を抽出する，という 2 つのステップが必要となる．本研究では，ステップ 1 は論文間の引用情報を，ステップ 2 は論文表題の解析技術を，それぞれ用いる．論文間の引用情報とは，論文間の引用･被引用関係だけでなく，ある論文が引用論文をどのような理由で引用しているのか(引用タイプ)も含めた情報のことである[1]．ステップ 1 では，キーワード検索により特定分野の論文を収集するが，この時，同時に引用情報も考慮することで，キーワードを含んでいない当該分野の論文も収集する．ステップ 2 では，収集された論文の表題から要素技術に関する情報を抽出する．多くの論文表題には「A に基づいた」や「B を用いた」などの表現が含まれる．この A や B は，ある技術を実現するための要素技術を示す用語であると考えられる．そこで，論文表題を解析し，要素技術を示す用語を抽出する．用語を抽出した論文の著作年を X 軸に，抽出された用語を Y 軸にとることで，ある分野の動向を示すグラフを作成することができる．

3.2 システム動作例

図 2 にシステムの動作例を示す．図は，「形態素解析」という用語をシステムに入力した時の解析結果を示している．図において，左端に「形態素解析」の要素技術名が列挙してあり，その用語が論文表題中で使われた年が，各技術の右側に示してある．例えば図 2 の「コスト最小法」の場合，この用語を論文表題に含んだ形態素解析に関する論文が 1987 年に 1 件，1993 年に 2 件発表されている．これらは図中で「●」として表示されており，その間が直線で結ばれている．ユーザが●上にカーソルを重ねると，その論

文の書誌情報がポップアップ表示される. 図では,「コスト最小法」(一番右端の●)にカーソルを重ねた時のポップアップ表示として「小松, コスト最小法に基づく逐次確定型･形態素解析, 1993」が例示されている.

さて, 図 2 において要素技術として提示されている用語をユーザがクリックすると, その要素技術が他にどのような分野で利用されているのかが一覧表示される. 図 3 は, 図 2 中の“hmm”(隠れマルコフモデル)をクリックした結果を示している. この図から, 1988 年には“speech recognition”(音声認識)の分野で, また, 2001 年には“summarization”(要約)の分野でそれぞれ利用されていることがわかる.

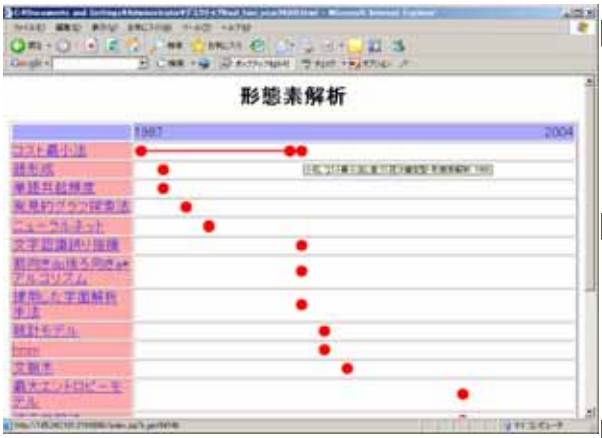

図 2 ある分野で使われている要素技術の一覧表示

4. おわりに

本稿では, 特許と論文データベースを統合し, 技術動向分析を行うツールを紹介した. 今後は, まず論文用語から特許用語の変換手法の定量的な評価を行う. 次に技術動向分析ツールを論文だけでなく特許にも拡張する.

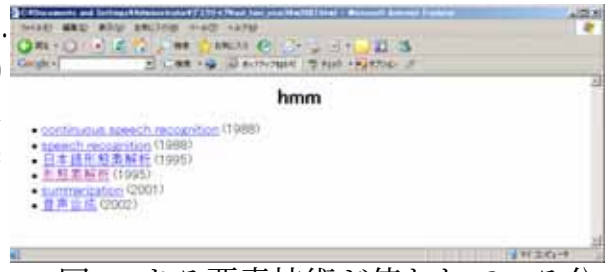

図 3 ある要素技術が使われている分野の一覧表示

5. 謝辞

本研究で用いた特許データ(1993 年～2002 年の特許公開公報)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.

6. 参考文献

[1]Nanba H., Abekawa T., Okumura M. and Saito S., “Bilingual PRESRI: Integration of Multiple Research Paper Databases” Proc. of RIAO 2004, pp.195-211, 2004.

[2]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, pp.743-746, 2006.

[3]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.

[4]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.

[5]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.