2005－ＮＬ－168（４）
2005／7／22

# 特許，論文データベースを統合した検索環境の構築

安善 奈津美[1] 難波 英嗣[2] 相沢 輝昭[2] 奥村 学[3]

1 広島市立大学 情報科学研究科 〒731-3194 広島県広島市安佐南区大塚東 3-4-1

2 広島市立大学 情報科学部 〒731-3194 広島県広島市安佐南区大塚東 3-4-1

3 東京工業大学 精密工学研究所 〒226-8503 神奈川県横浜市緑区長津田町 4259

E-mail: 1,2 {anzen, nanba, aizawa}@its.hiroshima-cu.ac.jp, 3 oku@pi.titech.ac.jp

**あらまし** 特許と論文間の引用関係を用いて両データベースを統合することにより，異なるジャンルの文書を横断的に検索するシステムの構築を目指し，本研究では特許中に引用されている特許と論文情報を自動的に抽出する手法を提案する．引用特許についてはパターンマッチングによる特許番号の抽出，引用論文については引用論文記述と共に出現する手掛かり語句を利用して書誌情報の記述を含む文の抽出を行う．提案手法の有効性を確認するための実験を行った結果，引用特許抽出実験では 99%の精度が，また，引用論文抽出実験では，精度 91%，再現率 86%が得られた．

**キーワード** 引用関係，情報検索，無効資料調査，論文，特許

# Construction of a Cross-genre Retrieval Environment by Integrating a Patent and a Research Paper Database

Natsumi ANZEN[1] Hidetsugu NANBA[2] Teruaki AIZAWA[2] and Manabu OKUMURA[3]

1,2 Hiroshima City University 3-4-1, Ozuka-higashi, Asaminami-ku, Hiroshima 731-3194 Japan

3 Tokyo Institute of Technology 4259, Nagatsuta, Yokohama 226-8503 Japan

E-mail: 1,2 {anzen, nanba, aizawa}@its.hiroshima-cu.ac.jp, 3 oku@pi.titech.ac.jp

**Abstract** In this paper, we propose a method to detect cited literatures in patents, as a first step towards construction of a cross-genre retrieval environment by analyzing citation relationships between a patent and a research paper databases. We detect cited patents by using simple regular expressions while we detect cited papers by combining several cue phrases. From the results of our examinations, we obtained 99% precision for the detection of cited patents, and 91% precision and 86% recall for that of cited papers.

**Keyword** Citation Relationships, Information Retrieval, Invalidity Search, Research Paper, Patent

## 1. はじめに

本研究では，特許の無効資料調査を支援するシステムの開発を行う．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のことで，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．加えて，サーチャーが審査官による審査を経た出願技術を再調査し，競合する他者の権利を無効化するために行う社内調査などもまた，無効資料調査と呼ぶ．こうした調査には，特許や論文など様々な情報が対象になる．NTCIR4，5 における特許検索タスクでは，特許を対象にした無効資料調査課題が設定されている[2]．これに対し，本研究では，特許だけでなく論文にも対象を拡大した無効資料調査を支援するシステムの開発を目指す．

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，一般的な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．また，

一般性の高い用語をキーワードに用いた場合，目的とする分野以外のものも含む，膨大な件数の特許が検索される可能性もある．さらに，関連する特許や論文を網羅的に調べるには，複数の言語を対象にしなければならない．海外の特許や論文を検索するには，日本語キーワードとは別にキーワードを用意し，海外の特許や論文データベースを検索する必要があるが，言語の違いは単語ベースで関連文書を探す上で非常に大きな障壁となり，さらに検索作業を煩雑にするものと考えられる．従って，論文と特許を横断的に検索するには，キーワード以外の検索方法も検討する必要がある．

一般に，特許や論文では，それぞれ関連研究や関連特許を引用する慣習がある．近年では，特許中で関連論文を，逆に論文において関連特許が引用されるケースも増えている．そこで，このような文書間の引用関係を辿れば，論文や特許と関連する文書を集めることができる．さらに，海外の特許や論文データベースとも統合すれば，引用関係を辿って，海外の特許や論文も収集が可能になる．本研究では，特許と論文間の引用関係の解析を解析することで，特許，論文データベースを統合する．

これらのデータベースが統合できれば，無効資料調査以外の様々な目的に利用できると考えられる．例えば，特許，論文間の引用関係に，種々の引用分析技術を適用することで，技術動向調査の支援を実現できる可能性がある．また，特許にあまり馴染みのない研究者が出願書類を作成する際，関連特許の調査にも利用できると思われる．これまでにNTCIR3の特許タスクでは，新聞記事を入力クエリとして，関連する特許を検索するジャンル横断検索が課題のひとつとして設定されていたが［3］，本研究では，特許と論文間の引用に着目したジャンル横断検索の実現を目指す．

本稿の構成は以下の通りである．2節では特許と論文における引用関係と抽出，3節では引用論文抽出に用いる手掛かり語について，4節では本手法の評価実験とその結果について述べる．

## 2. 特許，論文中の引用関係の抽出

### 2.1. 引用関係の解析

引用関係を用いて特許と論文データベースを統合するには，次の4つの手順が必要である．

(1) 論文中の特許への引用箇所の抽出
(2) 特許間の引用関係の抽出
(3) 特許中の論文への引用個所の抽出
(4) 同じ内容の特許と論文の対応付け

本研究では，引用論文データベースPRESRI[1]と特許データベースとの統合を試みる．PRESRIとは，Web上のPostscript及びPDF形式の日英論文データ約9万件を収集して構築され，引用関係を解析して図示することを可能とした，論文間の関係を利用したデータベースである［3］．上記の4つの手順のうち，(1) の論文間の引用解析については，このPRESRIの技術で対応可能である．従って，本研究では手順 (2)，(3) を扱う．

手順 (2)，(3) において，2003年7月以前は，特許中での引用文献記述は従来技術という項に記載されていたが，その後，記述形式が変更され，特許文献，非特許文献という項目で列挙されることになった．これにより，引用文献の記載位置が計算機でも容易に判別できるようなったため，最近の特許に関しては，手順 (1) 同様，PRESRIの技術で対応可能である．しかし，変更以前の引用文献記述については，新たな抽出方法が必要である．

変更以前の手順 (2) については，特許は付与された個別の識別番号を用いて表記されるので，簡単なパターンマッチングで実現できる．一方，手順 (3) については，特許中での論文の引用形式は様々であり，形式ごとに人手で抽出規則を作成するのは手間がかかり，実現は容易ではない．しかし，特許中で抽出すべき被引用論文の書誌情報を同定することは，人間にとってそれほど大変な作業ではない．そこで，特許中で抽出すべき論文の書誌情報に人手でタグを付与し，抽出規則の自動獲得を試みる．

以下では，手順 (2) (3) について述べる．

### 2.2. 特許中の引用特許の抽出

**引用特許の抽出**

論文の関連研究，参考文献の書誌情報は論文の最後にまとめて記述されるが，特許中ではそのような関連文書は，従来技術の項に記述されるのが一般的である．特許中では，他の特許を引用する際，出願番号や公開番号，登録番号が記述される．各特許には個別の番号が付与されるため，番号さえ分かれば，発明の名称や発明者等が特定できる．また，番号の表記方法には揺れが少ないので，簡単なパターンマッチングで抽出することが出来る．

**引用特許の記述**

特許の記述には以下のものがある．

**・出願公開公報**

通常の国内出願特許が掲載される．公開番号は，「公開された年」，「出願の種類」，「その年の出願順序」の組み合わせによって決まる．

例：２００４年特許出願公開１２３４５号
（特開２００４－１２３４５）

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

・**特許掲載公報**

審査を通り，特許された発明だけが掲載される．特許番号は，特許された順の連続番号で与えられる．

例：特許第１２３４号

・**国際公開**

国際出願した際の公開公報．国際出願番号は，公開された年，公開の際の連続番号によって決まる．

例：WO２００４/１２３４

なお，国際公開された国際特許出願のうち日本を指定国に含むものは，公表公報または再公表公報が発行される．この際に与えられる番号は公開公報と同様の形式である．

・**海外特許**

特許中で様々な国の特許が引用されることがある．その表記は，国名と番号の組み合わせで記述される．

例：アメリカ特許第 49238 号(ＵＳＰ49238)

欧州特許第９１３４号

## 2.3. 特許中の引用論文の抽出

### 2.3.1. 引用論文の記述

特許中の引用論文もまた引用特許と同様に，従来の技術の項に記述される．論文の書誌情報の記述の例を図１に示す．

1. 従来のオンライン文字認識方法の一例が，1995 年，若原徹他，ストローク単位のアフィン変換を用いたオンライン手書き漢字認識（電子情報通信学会技術報告書 PRU95-111 pp. 49-54）に記載されている．
2. この種のマルチタスクシステムとしては，電子情報通信学会技術研究報告ＣＱ９６－１７として発表されている．
3. このような推論方式を設計の支援に応用した例がある(人工知能学会誌 1992. 7)．
4. 例えば，語彙機能文法 (Kaplan, Ronald M., and Bresnan, Joan. (1982). "Lexical Functional Grammar:A formal system for grammatical representation." In Joan Bresnan, editor, The Mental Representation of Grammatical Relations, pages 173-281. MIT Press, Cambridge, Mass) を参照されたい．

図１　引用論文記述例

図１に示す４つの文はそれぞれ異なる「従来の技術」から抜粋したものである．このうち，1 では，著作年，著者名，タイトル，掲載誌，掲載ページが記述されているが，2，3 では著者名やタイトルが記述されていない．この他，海外の論文を引用する場合も４のように，国内の論文と同様の形式で引用されるが，日本語（カタカナ）で書き直される場合もある．

特許と違い，論文には個別の番号が与えられてはいない．よって，ある論文と別の論文が同一のものであるかどうかを判断するには，タイトル，著者名，掲載誌（巻，号，頁），著作年といった出来得る限り多くの項目を照合する必要がある．従って，特許中に記述されている引用論文の項目は，可能な限り全て抽出できれば望ましい．しかし,前述した通り，引用論文の記述方法は多様であり，引用部分の記述を抽出する規則を人手で作成するのは非常に困難である．そこで本研究では，特許から引用論文記述に該当する部分を絞り込み，書誌情報の抽出を行う．

### 2.3.2. 引用論文の抽出

図２は引用論文を含む「従来の技術」の項の全文である．なお，説明のため，各文頭に (1) ～ (3) の文番号を付与している．

【従来の技術】(1) 従来，この種のオンライン文字認識装置は，高精度に文字を認識するために，入力パタンと標準パタンとの間でストロークまたは特徴点間の対応付けを行う必要がある．(2) この対応付けでは，手書きの変動に対しても頑健に認識するために，筆順や画数の制約を課したり，標準パタンを入力パタンへの重なりが最大となるように適応的整形を加えたりしている．(3)従来のオンライン文字認識方法の一例が，1995 年，若原徹他，ストローク単位のアフィン変換を用いたオンライン手書き漢字認識(電子情報通信学会技術報告書 PRU95-111pp. 49-54)に記載されている．

図２　従来技術項目の例

この３文の中から，タイトルや著者名の項目を抽出するのも一つの方法であるが，本研究では，まず，引用論文の書誌情報が記述されている文 (3) を特定し，次に，文 (3) から各項目の抽出を行う．この２段階の処理のうち，今回は前者について取り組む．後者については，先行研究[1]を参考にする予定である．文 (3) の抽出には，書誌情報記述の際によく使われる語句を手掛かりに判定を行う．例えば図２の例の場合，「pp」や「に記載」等が手掛かりとして利用できる．

## 3. 手掛かり語の抽出

引用論文記述を含む文の抽出には，書誌情報を記述する際に使用される語句の有無を判定に用いる．例えば図１の４に示す例では，書誌情報の前に「例えば,」といった表現がある．また，書誌情報の後に「参照」

という表現もある．これらの表現は書誌情報を含む文の抽出に有効な手掛かりになると考えられる．そこで本研究では，この様な書誌情報記述と共に出現する回数の多い語を手掛かり語とし，その組み合わせにより論文情報の抽出を試みる．

いくつかの「従来の技術」を分析した結果，手掛かり語には，以下に示す3種類があることがわかった．

- ポジティブな手掛かり語
  - ➢ 前後　（例）「例えば」「記載」「参照」
  - ➢ 中　（例）「pp.」「Vol.」
- ネガティブな手掛かり語
  （例）「新聞」

「前後」手掛かり語は，引用記述部分の前後に出現する語句である．「中」手掛かり語は引用記述中に出現する語句である．前後手掛かり語と中手掛かり語を合わせてポジティブな手掛かり語と呼ぶ．これに対し，論文を引用するときに現れない表現をネガティブな手掛かり語と呼ぶ．例えば，論文と新聞が一文中で同時に引用されることは滅多にないため，「新聞」という表現が現れた場合，その文中には論文の書誌情報は記述されていないと判断してよい．しかし，このような手掛かり語を人手で網羅的に収集するのは容易ではない．この様な場合，対象テキスト集合に n-gram 等を適用して，手掛かり語の候補を収集するのが一般的であるが，特許中で論文が引用される割合はそれほど高くないので，手掛かり語候補の中に多くのノイズが混じってしまう．そこで本研究では，次に述べる手順で効率的な手掛かり語の収集を試みる．

まず2種類のコーパスを用意する．ひとつは PRESRI が保持する約9万件の論文書誌の出典情報コーパスである．もうひとつは特許データベースから任意に抽出した「従来の技術」中で論文の記述がある文に人手でタグを付与したコーパスである．前者には，「pp.」や“Proceedings of”といった中手掛かり語が多く含まれると考えられるので，このコーパスに n-gram を適用すれば，中手掛かり語を効率的に収集できる．これらの各手掛かり語が「従来の技術」の文中に出現するか否かを素性と考え，後者のコーパスを訓練用データとして，次節で述べる方法で機械学習を行い，引用論文の書誌情報記述を含む文の抽出器を獲得する．この抽出器を用いて「従来の技術」から文を抽出する．これらに n-gram を適用し，前後手掛かり語及び中手掛かり語を選定する．また，抽出しなかった文を収集し，n-gram を適応してネガティブな手掛かり語を選定する．新しく得られた手掛かり語を用いて，再度機械学習を行い，文抽出器を獲得する．この様な手順を繰り返すことで，使用する手掛かり語を増やしていく．以上の手順をまとめると，以下のようになる．

(1) 論文の出典情報集合から頻出する語句を抽出し，それらを手掛かり語とする．
(2) 収集した手掛かり語を用いて，特許から引用論文の書誌情報を含むと判断した文を収集する．
(3) (2) で収集した文から頻出する語句を抽出し，手掛かり語に追加する．
(4) (3) で新たな手掛かり語の追加があれば，(2) に戻る．

最終的に前後手掛かり語を 14 語，中手掛かり語を 22 語，ネガティブな手掛かり語を 2 語得た．

## 4. 評価実験

### 4.1. 評価方法

提案手法の有効性を確認するため，2 種類の評価実験を行った．

**引用特許抽出実験**

特許中の引用特許の抽出には，特許公開公報，特許掲載公報に掲載されているものを国内特許とし，国際公開，その他海外で取得された特許を海外特許とし，それぞれ抽出規則を作成した．

実験に使用したデータは，特許公開公報 1993 年～2002 年の物理学：計算，係数：電気デジタルデータ処理分野から任意に選択した 25,000 件の従来技術項目である．これらの引用文献の記述箇所に人手でタグを付与した．国内特許のタグが付与されたものは 4,975 件，海外特許は 450 件であった．

評価は，以下に示す精度と再現率で行う．

- 精度

$$\frac{\text{規則を用いて抽出できた正解データ数}}{\text{規則を用いて抽出したデータ数}}$$

- 再現率

$$\frac{\text{規則を用いて抽出できた正解データ数}}{\text{全正解データ数}}$$

**引用論文抽出実験**

3 節で述べた手法で手掛かり語の選定を行い，機械学習により抽出規則を得た．使用した学習器は Support Vector Machine，カーネルは 2 次の多項式関数である．訓練に用いた特許データは 32,537 文であり，そのうち 1,186 文に論文タグが付与された．また，評価用データは 9,5362 文あり，正解数は 290 文であった．評価尺

度は精度と再現率を用いる．精度と再現率は引用特許抽出と同式を用いる．

提案手法と比較するため，2種類のベースラインを用意する．ひとつは前後手掛かり語のみを使用して学習を行ったもので，もうひとつは中手掛かり語のみを使用したものである．

### 4.2. 結果

#### 4.2.1. 引用特許抽出結果

結果を表1に示す．

表1　引用特許抽出実験の精度と再現率

| | 精度 | 再現率 |
|---|---|---|
| 国内特許 | 0.995 (4951/4972) | 0.995 (4951/4975) |
| 海外特許 | 0.957 (403/413) | 0.895 (403/450) |

表1に示す通り，国内特許はほぼ100%に近い精度と再現率が得られた．国際特許は国内特許には至らないものの，高い精度と再現率が得られた．

#### 4.2.2. 引用論文抽出結果

結果を表2に示す．

表2　引用論文抽出実験の精度と再現率

| 手法 | 精度 | 再現率 |
|---|---|---|
| 手掛かり語(前後) | 0.079 (257/3222) | 0.886 (257/290) |
| 手掛かり語 (中) | 0.309 (252/815) | 0.869 (252/290) |
| 提案手法 | 0.916 (252/275) | 0.869 (252/290) |

再現率はどの手法もほとんど差がないが，精度は提案手法は二つのベースラインと比べ，非常に高い値が得られた．

### 4.3. 考察

**引用特許抽出**

引用特許の抽出は特許番号のパターンマッチングで行っているので，特許番号が表記されていない場合や国内外の別が明記されていないものは抽出に失敗している（図3）．

・平成９年１月１６日出願〝２進桁上カット加算器〟
・コダック社出願番号第６１，４９６号

図3　引用特許の抽出失敗例

特許を引用する際，特許番号が記述されていないものはごく稀であるので，図3のような例は，本研究では処理対象外とする．特許出願をした国が明記されていない場合，特許全体を調べれば判明するかもしれないが，自動的に抽出するのは困難であるため，この様な場合も本研究では扱わないものとする．

**引用論文抽出**

引用論文記述のある一文の抽出では，手掛かり語をそれぞれ単独で用いた場合と比較して，提案手法の方が非常に高い精度を得ることが出来た．これは最適な手掛かり語の組み合わせが学習でき，過剰抽出が減少した結果だと考えられる．

抽出に失敗したものは，タイトルと著者名しか記述されていない場合であった．また，引用の記述形式が論文と似ている書籍や新聞記事からの引用などが誤って抽出されてしまった（図4）．

・従来，カラー画像処理装置における，階調表現に用いられる誤差拡散処理の例としては，Floyd-Steinberg による文献〝An Adaptive for Spatial Grey Scale〟が知られている．
・例えば，１９９９年５月８日付け日経産業新聞の記事「ペンで家電を操作　玉川大とソニーＣＳＬ紙をなぞると作動」は，次のような文章で，簡易入力装置を紹介している．

図4　引用論文の書誌情報を含む文の抽出失敗例

しかし，これらの抽出失敗については，次の論文情報の同定処理の段階において，情報不足で判定できないため，影響はない．

## 5. おわりに

本研究では引用関係を利用した特許と論文データベースの統合を実現するため，特許中で引用される関連文献の抽出を行った．引用特許の抽出に関しては人手で抽出規則を作成し，国内特許はほぼ100%，国際特許は国内特許には至らないものの，高い精度と再現率が得られた．

引用論文抽出については，引用論文記述と共に出現する手掛かり語を収集し，これを用いて引用論文記述を含む文の抽出を行った．その結果，精度91.6%，再現率86.9%が得られた．

## 今後の課題

引用論文抽出においては，今回抽出した引用論文記述を含む文からタイトルや著者名といった書誌情報を，阿辺川らの手法を参考に抜き出し，これらをPRESRIの論文データと照合，比較する処理が必要である．その後，2節で述べた4つの手順のうち，「(4) 同じ内容の特許と論文の対応付け」に取り組む予定である．

## 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，ウェブ・アンド・ゲノム・インフォマティクスの新森昭宏氏，デュオシステムズの宮原俊一氏，鈴木泰山氏に感謝致します．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．

## 文　献

[1] 阿辺川武，難波英嗣，高村大也，奥村学，“機械学習による科学技術論文からの書誌情報の自動抽出，”情報処理学会 自然言語処理研究会，NL-157，pp.83-90，2003.

[2] A. Fujii and T. Ishikawa, “Document Structure Analysis in Associative Patent Retrieval,” Working Notes of NTCIR-4, pp233-237, 2004.

[3] M. Iwayama, A. Fujii, N. Kando, and Y. Marukawa, “An empirical study on retrieval models for different document genres: patents and newspapers articles,” Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.251-258, 2003.

[4] 難波英嗣，阿辺川武，奥村学，齋藤豪，“Web 上のデータを中心とした複数論文データベースの統合，”言語処理学会 第 11 回年次大会，2005.