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 広島市立大学大学院　情報科学研究科
2 広島市立大学　情報科学部
3 東京工業大学　精密工学研究所

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．これには，民間（企業）の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含まれる．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[3]．これに対し本研究は，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がある．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．このような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，検索作業を煩雑にする．従って，国内および海外の特許や論文に関するデータベースを横断的に検索可能なシステムの開発が必要となる．

一般に，特許や論文では，それぞれ特許中で関連特許を，論文中で関連研究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を引用するケースも増えている．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．これによって無効資料調査に伴う煩雑な検索手順を軽減することが可能となる．

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[4]，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について，3 節では本研究での提案手法である，引用論文の書誌情報の抽出について，4 節では本手法の評価実験とその結果について述べる．5 節では本研究での提案手法を用いた検索システムの動作例を紹介する．

2．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箇所の抽出
（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を収集して構築され，論文間の引用関係を解析して図示することを可能としたデータベースである[1][6]．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．続いて，手順 (2)，(3) での特許文書中の引用関係の抽出について説明する．

従来，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，その記述方法は 3 節図 1 に見るように様々であった．ところが 2003 年 7 月以降，記述形式が変更され，「背景技術」という項目で，特許文献，非特許文献に分けて文献情報のみが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] http://www.presri.com/

（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．以下,「従来技術」からの文献情報の自動抽出について考える．

変更以前の従来技術に対する手順（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［2］．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．3 節では，この手順（3）での特許中の引用論文抽出について述べる．

３．特許中の引用論文の抽出

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 と 3 では著者名やタイトルが記述されていない．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．

なお，引用論文を含む一文の抽出については，手掛かり語を素性とし，機械学習により抽出規則を獲得する手法が実現できている[2]．そこで本論文では，次の段階である書誌情報の各構成要素の抽出について扱う．

3.2 書誌情報の抽出

引用論文表記の多くは，表題，著者名，出典，著作年の組み合わせで構成されている．そこで，抽出にはそれぞれの特徴を利用する．

表題　：句点やピリオドが含まれず，平仮名，カタカナ，漢字等が連続する．
著者名：数文字おきに読点が出現する．
出典　：Vol,巻,号といった語句が出現する．
掲載頁：「pp」の後に数字が続く．
　　数字の後に「ページ」「頁」が出現する．
著作年：「19」または「200」から始まる４桁の数.

以上の特徴を考慮し，ある形態素の前後にどんな種類の形態素があるか，あるいは手掛かり語があるかどうかで引用論文の書誌情報が抽出できると考えられる．本研究では以上で挙げた特徴を素性とし，機械学習により抽出規則の獲得を行う．機械学習を用いた書誌情報抽出として，阿辺川らの先行研究がある[1]．阿辺川らは論文末尾の参考文献の書誌情報を抽出する際，学習モデルの作成及び各書誌情報タグの付与にYamCha[2]を用いている．本研究でも阿辺川らと同様にYamChaを用いて書誌情報の抽出を行う．しかしながら，本研究で対象としているのは特許中の引用論文記述である．そのため，論文末尾の参考文献とは異なり，次のような特徴が挙げられる．

- 書誌情報以外の文字が出現する
- 出現する構成要素の種類と順番が定まらない
- 一文中に複数の論文が引用される場合がある
- 日本語，英語が併記される場合がある

阿辺川らの研究では文字単位で同定を行っているが，本研究では，タグの付与は形態素単位で行う.本研究で使用するタグの一覧を表１に示す.

続いて，本研究における YamCha でのタグ付与過程について図２を用いて説明する．図２は入力が特許文書の従来技術である，“ 例えば，情報処理学会第 42 回全国大会（平成３年）にて藤原秀人他によって発表された論文「監視用高速画像処理装置」に示されている ” という一文の解析である．文字列が入力されると，形態素ごとに著者名，

[2] http://www.chasen.org/~taku/software/yamcha/

表題，出典といった判定を行い，タグが付与されていく．図 2 の左から 2 列目以降が素性であり，最右列が解析結果として付与されたタグである．素性は，入力文字列自身とその品詞，引用論文記述を含む文の抽出に利用した手掛かり語（例：「巻」「号」「論文誌」）22 種［2］である．左端の列は判定中の形態素からの距離を表している．ここで「３」という形態素を判定する場合，「３」という形態素自身とその品詞，そして「３」の前後の形態素とその品詞，さらに「３」より前にどのようなタグが付与されたかという判定結果も素性として加えられ，これらの情報を用いて解析が行われる(図 2 太枠部分)．この判定に用いる素性の範囲をウィンドウサイズと呼ぶ．事前に行った調査の結果から，本研究ではウィンドウサイズは５としている．

表 1 書誌情報タグ

	英語	日本語
表題	＜E-TITLE＞	＜J-TITLE＞
著者名	＜E-AUTHORS＞	＜J-AUTHORS＞
出典	＜E-SOURCE＞	＜J-SOURCE＞
頁	＜PAGE＞	
著作年	＜DATE＞	
他	＜OTHER＞	

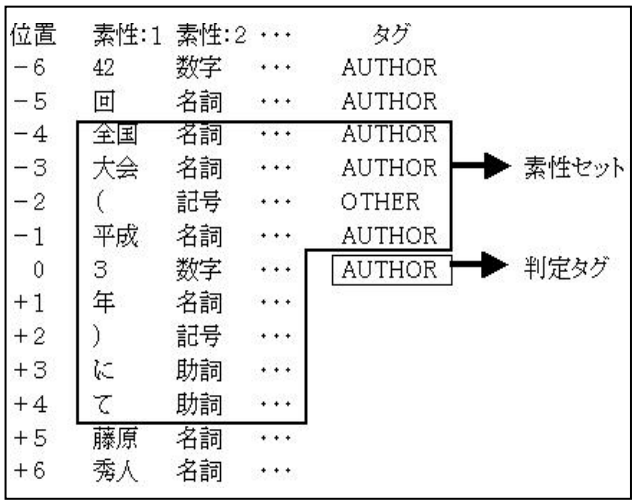

図 2 YamCha のタグ付与過程

本研究では，YamCha以外にCRF++[3]を使用し，比較を行う．CRF++はCRFを用いたチャンキングツールであり，少ないコーパスで効果的に学習することが可能であると言われており，品詞付与，固有表現抽出などの分野で高い精度が得られている[5]．なお，CRF++ではウィンドウサイズを 2 としている．

[3] http://www.chasen.org/~taku/software/CRF++/

４．評価実験

提案手法の有効性を確認するため，評価実験を行った．使用したデータは引用特許抽出実験，引用論文抽出実験と同様，特許文書従来技術のうち，特許公開公報（1993 年～2002 年）国際特許分類 G06F に属する特許文書の従来技術項目のうち，引用論文記述を含む 3,000 文であり，あらかじめ書誌情報タグが人手で付与されている．評価尺度には以下の式で示す精度と再現率を用いた．

・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

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

評価単位については，提案手法では形態素単位でタグ付与を行っており，同じタグを持つ形態素を連結させて文字列を作成しているので，同じタグを有する文字列単位で正解データの文字列と比較を行う．その際，漢字，平仮名，数字，アルファベット以外の記号だけが異なる場合は正解しているものとみなす．

評価実験の結果を表 2 に示す．表 2 からわかる通り，書誌情報抽出では，YamCha よりも CRF++ を使用した方が良い結果を得ることが出来た．また，各書誌情報の項目を比較すると，出典，著者名，表題の精度，再現率の値が低い結果となった．これは，全く抽出できなかったものもあるが，著者名を出典扱いしてしまうなど，互いに誤って抽出してしまっていることが原因と考えられる．今後はこれらの誤抽出を減らしていく処理が必要である．その対応策の一つとして，出典に含まれる学会誌名のリストを抽出の際に利用することを検討中である．

表 2 書誌情報の抽出結果

		YamCha		CRF++	
		精度	再現率	精度	再現率
英語	著者名	0.720	0.787	0.872	0.857
	出典	0.752	0.787	0.805	0.799
	表題	0.763	0.901	0.746	0.903
日本語	著者名	0.742	0.715	0.885	0.765
	出典	0.733	0.662	0.834	0.736
	表題	0.868	0.880	0.848	0.881
頁		0.941	0.932	0.973	0.973
日付		0.897	0.897	0.932	0.921
平均		0.802	0.822	0.862	0.854

５．動作例

本研究の提案手法を用いて構築した検索システムの動作例を図 3，図 4 に示す．引用関係を利用して特許と論文データベースを統合することで，異なるジャンルの文献である特許と論文の横断的な検索が可能である(図 3)．また，検索結果の引用関係を図示して提示することもできる(図 4)．

図 3　キーワード検索画面

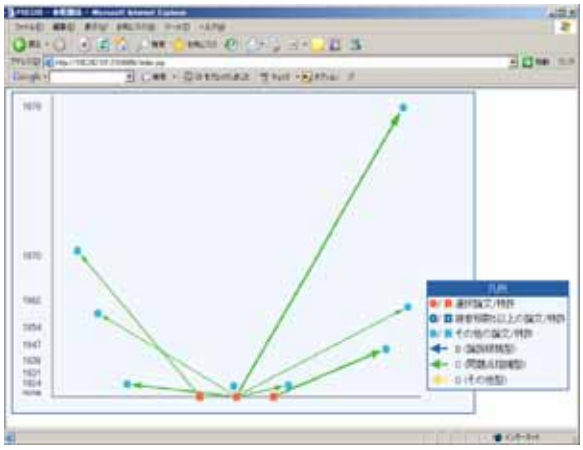

図 4　特許，論文間の引用関係のグラフ表示

図 4 において，■は特許を，●は論文を示している．さらに，「論説根拠型」，「問題点指摘型」，「その他」で表される，技術文献間の引用関係の種類 [6] が，矢印の色の違いにより表示される．なお，グラフ中の■や●の座標は，縦軸は著作年であり，横軸はランダムな数値を割り当てている．では引用関係を示すだけではなく，■や●をクリックすることにより，著者名，表題等の，特許や論文の詳細情報を表示することができる．このように，あるトピックに関連する複数の特許や論文をグラフとして提示することで，そのトピックに関する研究や技術動向の，直感的・視覚的な理解が可能という利点を持つ．

６．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論文における書誌情報の構成要素の抽出を行った．書誌情報の各構成要素の特徴を利用し，機械学習で抽出規則を獲得した．実験の結果，本研究の提案手法で，特許中の引用論文の書誌情報が実用的な精度で抽出できることがわかった．

今後の課題

今後の課題としては，書誌情報抽出の精度向上の他，一文中に複数論文が引用されている場合に必要な論文毎の書誌情報の切り分け，出典の省略への対応などが挙げられる．次の段階としては，本研究で抽出を行った書誌情報を PRESRI の論文データと照合，比較する処理が必要である．その際，抽出段階での解析ミスが，どの程度照合段階に影響するのかについても調査する必要がある．

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，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，ピコラボの鈴木泰山氏に感謝致します．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

[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，”情報処理学会研究報告，NL-157，pp.83-90，2003.

[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 ”，情報処理学会研究報告 NL-168，pp.21-26，2005.

[3] A. Fujii and T. Ishikawa, “ Document Structure Analysis in Associative Patent Retrieval, ” Working Notes of NTCIR-4, pp.233-237, 2004.

[4] 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 SIGIR ' 03, pp.251-258, 2003.

[5] 工藤拓，山本薫，松本裕治，“ Conditional Random Fields を用いた日本語形態素解析 ” 情報処理学会研究報告 NL-161，pp.89-96，2004.

[6] 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，pp.711-714，2005.