米国特許データベースからの引用文献情報の抽出

小栗　佑実子[1]　　難波　英嗣[2]
1広島市立大学情報科学部
2広島市立大学情報科学部

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．

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．本研究では，無効資料調査に伴う煩雑な検索手順を軽減するため，特許と論文間の引用関係の解析を解析することで，特許，論文データベースの統合を行う．これまでに，国内の公開広報を対象に引用文献を抽出する手法が提案されているので[2]，本研究ではその手法を基に，米国特許から引用文献情報を抽出する．

本稿の構成は，以下に示す通りである．２節で特許と論文における引用関係について，３節で本研究での提案手法，すなわち特許と論文データベースの統合を行うために必要な引用関係の抽出方法について述べる．４節では，本研究で提案した手法について評価実験と結果，それに基づき考察を行う．

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][4]．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．また，手順（２）に関しても，人手で抽出された米国特許中の引用特許データを用いる．よって，本研究では手順（３）を扱う．

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，実現は容易でない．また，引用論文記述が従来技術という項目に記載されている国内特許と異なり，米国特許では論文が引用されている項目が様々であるため，項目名から論文の引用個所を絞り込むことが出来ない．よって，まず引用論文記述が含まれている文を抽出し，そこから，論文表題や著者名などの書誌情報の抽出を行う．

３．特許中の引用論文の抽出

３．１　引用論文の記述

特許中の引用論文の書誌情報の記述例を図１に示す．図１示す 4 つの文はそれぞれ異なる特許中から抜粋したものである．このうち，1 では著者名，表題，掲載誌，掲載頁，著作年が記述されているが，2 では表題，著者名が記述されていない．この他，海外の論文を引用する場合は 3 のように

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

論文の書かれた言語のまま引用される場合もあるが，英語に訳されて引用される場合もある．

1. For example, selective catalytic reduction of nitrogen oxides with vanadium-impregnated monolith catalysts is accelerated by the introduction of large pores to the monolith, as taught by Beeckman and Hegedus in "Design of Monolith Catalysts for Power Plant Emission Control, " in Industrial & Engineering Chemistry Research, Volume 29, pp. 969-978, 1991.
2. Journal Solid State Circuits, vol. 25, no. 4, pp. 1028-1031, August 1990, describes bipolar circuits that include integrated inductors.
3. This experiment was conducted according to Kogyo Kagaku Zasshi, Vol. 72, p. 2081 (1969). 19 ml of toluene, 0.06 mmole of acetylacetone nickel and 0.75 mmole of titanium tetrachloride were charged in a 100 ml autoclave equipped with stirrer and maintained under a nitrogen atmosphere.

図 1：引用論文記述例

特許と異なり，論文には個別の番号が与えられていない．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，表題，著者名，出典（掲載誌名，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．よって，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．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[2]に基づき行う．

３．２　引用論文の抽出

引用論文記述が従来技術という項目に記載されている国内特許と異なり，米国特許では論文が引用されている項目が様々であるため，項目名から論文の引用個所を絞り込むことが出来ない．そのため，本研究ではまず引用論文の書誌情報が記述されている一文を特定し，次にその一文から各項目の抽出を行う．引用論文の書誌情報が記述されている一文の抽出には，書誌情報記述の際によく使用されている語句を手掛かり語として判定を行う．手掛かり語の収集は，以下の手順で行う．

（1）国内特許からの引用論文抽出[2]で提案されている手掛かり語一覧から，英語のものを手掛かり語とする．
（2）手掛かり語を用いて，米国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 n-gram を抽出する．
（4）（3）で出力された n-gram の中に有効な手掛かり語があれば追加し，（2）に戻る．

以上の手順で最終的に得られた手掛かり語の一覧を表 1 に示す．

表 1：手掛かり語一覧

University，article，conference on，International，issue，Journal of，Letters，meeting，pages，pp.，Proceedings of，publication，published，Symposium，titled，transaction，Vol.，volume

これらの手掛かり語を用いて，引用論文記述を含んでいると考えられる文の抽出を行う．ここで，表 1 の手掛かり語がある文に出現すれば，その文中に論文記述があるという保証はない．そこで，どのような手掛かり語がどのような条件で出現したときに，その文に引用論文記述が含まれると判断するべきかを検討する必要がある．本研究では，「文中に手掛かり語がどのような組み合わせで出現するのか」や「どういった順番で手掛かり語が出現するか」といった情報を用いて引用論文記述を含んだ文の抽出を行う．これらの情報を素性として与え，機械学習により抽出規則を獲得する．機械学習を行う際に，表 1 の手掛かり語に加え機械学習に特化した手掛かり語を追加した．機械学習に使用した手掛かり語を表 2 に示す．

表 2：機械学習に使用する手掛かり語一覧

前後	article，issue，published，publication，titled，for example，described，presented，reported，reference
中	University，Conference on，International，Journal of，Letters，pp.，Proceedings of，Workshop，Symposium，Transaction，meeting，Vol.，no.，*:*，19**|20**，et al.

手掛かり語は，「前後」手掛かり語と「中」手掛かり語の 2 種類に分けられる．「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．

３．３　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　：「”」という記号が表題の前後に出現することが多い．
著者名：数字置きに「.」が出現する場合や「et al.」という単語が出現する．
出典　：「Vol.」，「No.」といった語句が出現する．
掲載頁：「pp.」のあとに数字が続く．
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，ある単語の前後にどんな種類の単語があるか，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と，単語の品詞を素性とし，機械学習により抽出規則の獲得を行う．単語の品詞付けには，TreeTagger[2]を用いる．

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，機械学習を用いて抽出規則の獲得を行っている．しかし，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．

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，タグの付与は単語単位で行う．本研究で使用するタグの一覧を表 3 に示す．

表 3：書誌情報タグ

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

また，阿辺川らの研究では，学習器に YamCha[3]を用いているが，本研究では，CRF++[4]を使用する．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．

４．評価実験

４．１　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獲得する．使用した学習器は Support Vector Machine(TinySVM[5])，カーネルは 2 次の多項式関数である．使用したデータは，米国特許庁で権利が認められた 993,490 特許（1993 年～2000 年）を用いる．この特許データから，表 1 の手掛かり語をひとつでも含む文を全て抽出し，そこから任意に選択した 10,310 文に対し，被験者が引用論文記述を含むかどうかの判定を行った．その結果，全部で 2,295 に引用論文記述が含まれていた．10,310 文のうち，9,375 文(含，正解 2,114 文)を訓練用に，残り 935 文(含，正解 181 文)を評価用に用いる．

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

実験の結果を表 4 に示す．

2 http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

3 http://www.chasen.org/~taku/software/yamcha/

4 http://www.chasen.org/~taku/software/CRF++/

5 http://chasen.org/~taku/software/TinySVM/

表 4：引用論文を含む一文の抽出結果

	精度（%）	再現率（%）
手掛かり語	22.27	100.00
提案手法	91.33	87.78

手掛かり語のみを使用し抽出を行う時に比べ，機械学習を行う提案手法では非常に高い精度が得られた．これは，素性の組み合わせが最適なものが学習でき，過剰抽出が減少した結果であると考えられる．

４．２　書誌情報抽出実験

機械学習を用いて，引用論文記述の書誌情報の構成要素の抽出を行った．使用したデータは引用論文抽出実験で被験者が正解と判定した 2,295 文に，表 3 に示す書誌情報タグを付与したものである．

評価尺度は，引用論文抽出実験と同一のものを用いる．また，評価単位については，提案手法では単語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．このとき，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 5 に示す．

表 5：書誌情報の抽出結果

	精度（%）	再現率（%）
E-AUTHORS	84.87	82.25
E-SOURCE	76.39	70.48
E-TITLE	76.64	66.94
PAGE	94.71	94.45
DATE	92.58	94.59
全体	85.02	82.06

表 5 の各書誌情報の項目を比較すると，出典，表題の精度，再現率の値が低い結果となった．これは，全く抽出出来なかったものもあるが，出典を表題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．その対応策として，出典に含まれる学会誌名や学術雑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．

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，引用論文の書誌情報抽出の精度向上や，一文中に複数の引用論文記述が出現する場合に必要な論文ごとの書誌情報の切り分けなどが挙げられる．また，次の段階として，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

本研究で用いた米国特許データは，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

[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] 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.

[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.