論文と特許を対象にした技術動向分析

第 7 回、第 8 回 NTCIR ワークショップ特許マイニングタスク

広島市立大学大学院情報科学研究科准教授　難波　英嗣

PROFILE

1996 年東京理科大学理工学部電気工学科卒業。2001 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了。同年、日本学術振興会特別研究員。2002 年東京工業大学精密工学研究所助手。同年、広島市立大学情報科学部講師。2010 年 4 月広島市立大学大学院情報科学研究科准教授、現在に至る。博士 (情報科学)。言語処理学会、情報処理学会、人工知能学会、ACL、ACM 各会員。

nanba@hiroshima-cu.ac.jp　TEL 082-830-1584

東京工業大学大学院情報理工学研究科准教授　藤井　敦

1998 年東京工業大学大学院博士課程修了。博士（工学）。筑波大学大学院准教授等を経て、2009 年より東京工業大学大学院情報理工学研究科准教授。自然言語処理、情報検索、Web マイニングの研究に従事。2009 年度より特許版産業日本語委員会委員。

TEL

株式会社日立製作所　中央研究所
東京工業大学精密工学研究所客員教授　岩山　真

1992 年 (株) 日立製作所入社。文書検索、自然言語処理の研究に従事。
2009 年度より特許版産業日本語委員会委員。

TEL

東京工業大学総合プロジェクト支援センター特任准教授　橋本　泰一

1997 年 東京工業大学工学部情報工学科卒業。2002 年 同大学大学院情報理工学研究科計算工学専攻博士課程修了。同年 同大学同研究科 助手。2006 年 同大学統合研究院 特任准教授。現在、自然言語処理、情報検索、テキストマイニングに関する研究に従事。情報処理学会、言語処理学会、科学技術社会論学会、研究・技術計画学会各会員。博士（工学）。2010 年度特許版産業日本語委員会委員。

TEL

1 はじめに

本稿では、特許や論文などの技術文献を対象に、ある分野でどのような研究や技術開発が行われているかを分析する技術について、国立情報学研究所 (NII) が主催する第 7 回および第 8 回 NTCIR ワークショップにおいて、筆者らが行った特許と論文を対象とした情報処理のためのテストコレクションの構築研究 [2, 3] について述べる。

近年、大学研究者自身が関連論文だけでなく関連特許について情報を検索したり、特許を出願したりする機会が増えており、2010 年 5 月に政府の知的財産戦略本部が発表した「知的財産推進計画 2010」においても、推進計画 2006、2007、2008 および 2009 に引き続き、大学研究における特許情報の重要性が謳われている。

特許と論文を検索するのは、大学研究者に限った話ではない。例えば、特許庁の審査官は出願された技術が特許権の取得に該当するかどうか判断するために、過去に同様の特許が出願されたり論文が発表されたりしていないか調査する。これは一般に先行技術調査と呼ばれている。この他に、サーチャーと呼ばれる専門の担当者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民間企業の社内で行われる無効資料調査でも、論文と特許が検索対象となる。

こうした状況を鑑み、特許と論文を対象にした検索や動向分析など、さまざまな目的に利用可能な言語処理技術の開発を最終目標とし、そのための第一歩として筆者らが位置づけているのが、特許マイニングタスクである。

2 特許マイニングタスクの概要

特許マイニングタスクの最終目標は、ある分野の特許と論文から、図 1 に示すような技術動向マップを自動的に作成することである。図は、論文と特許を、「要素技術」と「効果」という観点から分類し、技術動向マップとしてまとめたもので、このような技術動向マップを自動生成するツールは、1 章で述べた先行技術調査や無効資料調査の支援ツールとして利用できる。

	効果 1	効果 2	効果 3
要素技術 1	[論文 A] [特許 X]		[論文 B]
要素技術 2	[論文 C]		
要素技術 3		[特許 Y]	[特許 Z] [特許 W]

図 1　特定分野の特許と論文から生成される技術動向マップの例

このようなマップを自動的に生成するためには、以下の 2 つの手順が必要となる。

- (手順 1) ある分野の特許と論文を網羅的に収集する。
- (手順 2) 手順 1 で収集された特許と論文から要素技術と効果の対を抽出し、技術動向マップとしてまとめる。

これらの 2 つの手順について、特許マイニングタスクでは、以下の 2 つのサブタスクを設定している。

- 学術論文分類サブタスク
- 技術動向マップ作成サブタスク

以下、これらのサブタスクの概要を述べる。

学術論文分類サブタスク

このサブタスクでは論文抄録に、特許分類体系のひとつである国際特許分類 (International Patent Classification: IPC) のコードを自動的に付与する。IPC は、特許文献の技術内容によって上から順に「セクション」、「クラス」、「サブクラス」、「メイングループ」、「サブグループ」の 5 階層から構成・分類されており、国際特許分類第 6 版ではサブグループのレベルで約 50,000 [1] の IPC コードが存在する。本サブタスクでは、最下層の「サブグループ」レベルの IPC コードを論文抄録に付与することを目的とする。図 2 は日本語論文の例である。ここで、<TOPIC-ID> は論文の ID を、<TITLE> と <ABSTRACT> は論文表題と概要を、それぞれ示している。タスクの参加者は、図 2 のような入力が与えられると、対応する IPC コードを自動的に出力するシステムを構築することが求められる。

```
<TOPIC><TOPIC-ID>312</TOPIC-ID>
<TITLE> 二値画像用高速符号化 / 復号 LSI</TITLE>
<ABSTRACT> 二値画像データを高速で符号化，復号する LSI を開発した．参照ラインデータ上に「基準色変化点」を探すのと並行して，それを参照するランのイメージデータを生成する方式により，復号性能を向上させた．また，符号化時と復号時共に同じ方向にデータが流れるパイプライン構成とし，さらに主な回路は共通化する構成によって回路を簡略化した．</ABSTRACT>
</TOPIC>
```

図 2　学術論文分類サブタスクの入力例

1）NICIR-7 特許マイニングタスクでは、これらのうち、学術分野とは関連性の低い分野を除外した 30,885 の IPC コードを対象とした。

技術動向マップ作成サブタスク

このサブタスクは、要素技術とその効果を示す表現を、特許や論文から自動的に抽出することを目的とする。例えば「PM 磁束制御用コイルを設けて閉ループフィードバック制御を施すため、電力損失を最小化できる。」という文が入力されると、図３に示すように、要素技術と効果を示す個所に、それぞれ“TECHNOLOGY”および“EFFECT”タグを自動的に付与する。ここで、“EFFECT”タグの中には、さらに“ATTRIBUTE”と“VALUE”という２種類のタグが付与されている。技術の効果に関する表現は多様であり、そのすべてを処理対象とするのは、現在の言語処理技術では非常に困難である。このため、例えば、「処理速度 (ATTRIBUTE) が向上する (VALUE)」や「ノイズ (ATTRIBUTE) が減少する (VALUE)」のように、技術の効果が「属性 (ATTRIBUTE)」と「属性値 (VALUE)」の対で表現できるもののみを対象とする。近年の自然言語処理分野では、テキスト中に出現する属性と属性値の対の抽出が活発に研究されており、技術の蓄積が急速に進みつつある。特許や論文中の属性と属性値の対で表現可能な技術の効果に関する表現の抽出も、このような既存の技術の利用が期待できる。こうして、ある分野の論文と特許から、図３に示すような要素技術と効果の対が抽出できれば、図１に示すような技術動向マップの自動作成が実現可能になると考えられる。

PM 磁束制御用コイルを設けて <TECHNOLOGY> 閉ループフィードバック制御 </TECHNOLOGY> を施すため, <EFFECT><ATTRIBUTE> 電力損失 </ATTRIBUTE> を <VALUE> 最小化 </VALUE></EFFECT> できる.

図３　技術動向マップ作成サブタスクに用いるデータの一例 (VALUE タグ内が文字列の場合)

なお、このサブタスクでは、図４のように VALUE タグが付与される表現が数値となるものも対象としている。

<TECHNOLOGY>CRF</TECHNOLOGY> を用いた手法では, <EFFECT><VALUE>0.935 </VALUE> の <ATTRIBUTE> 精度 </ATTRIBUTE></EFFECT> が得られた.

図４　技術動向マップ作成サブタスクに用いるデータの一例 (VALUE タグ内が数値の場合)

図５　技術動向マップ作成サブタスクの出力例 (解析精度の推移)

もし、例えば「形態素解析」や「機械翻訳」などの特定分野の論文や特許から図４に示すような精度値が抽出できれば、精度値を縦軸に、論文の著作年や特許の出願年を横軸にとることにより、精度値の時間的な推移を示すグラフが描画できる。図５は、「形態素解析」に関する複数の論文から、実際に精度値を人手で抽出し、グラフにまとめたものである。このグラフから、形態素解析分野では、1994 年頃から精度が 95% 以上に達しており、この分野の技術が成熟しつつあるということがわかる。ここで、2000 年に精度が若干低下しているが、これは評価に用いるデータが異なるためである。本来ならば、評価用データや実験条件が違えば、評価値の直接的な比較はできないが、この分野への新規参入を検討している企業にとって、参入する余地があるかどうかの判断材料として利用するという目的であれば、図５のようなグラフは十分に有用である。なお、関連した研究に村田ら [1] のものがある。村田らは、自然言語処理分野の論文概要から、「精度表現」、「主要な分野」、「言語名 (その研究が対象としている言語)」、「組織・人名」を、情報抽出技術を用いて抽出し、表として出力している。

3 システムの動作例

２章で述べたテストコレクションを用いた評価結果の詳細については、文献 [2, 3] を参照されたい。本章では、このテストコレクションを用いて著者らが作成したシステムの動作例について説明する。

図６は、「音声認識」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６において、左側に「音声認識」の要素技術が列挙され、各技術の右側にその用語が使われている年が示される。例えば図６中にある要素技術「HMM」の場合、この用語を要素技術に用いた文献が 1999 年に発表されていることを示している。これらは図中で「●」として表示されており、ユーザが●上にカーソルを重ねることで、その文献の書誌情報がポップアップウィンドウ内に表示される。

また、図６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ことで、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を、年代順に一覧表示することができる。図７は、図６中の「HMM」をクリックした結果を示している。学術界では 1990 年代前半に画像認識 (地名認識) の分野で使われていた

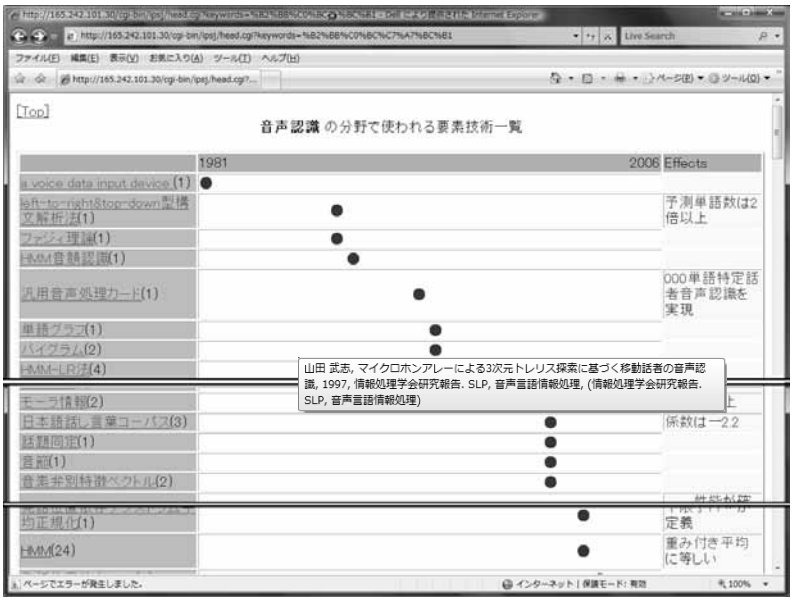

図６　「音声認識」で使われる要素技術と効果の一覧表示

図７　「HMM」を要素技術として用いている分野と各分野における効果の一覧表示

技術が 1990 年代後半に入ると動画像認識（ジェスチャ認識）の分野でも利用されていることが一覧表示される。

さらに、各要素技術の効果に関する情報が、各図の右端に表示される。図６では、「音声認識」の分野で「モーラ情報」の技術から「精度が向上」するという効果があることが分かる。また、図７では、様々な分野においてある要素技術にどのような効果があるのか一覧できる。

なお、ここで紹介したシステムは、要素技術とその効果を出力しているが、図１とは少し異なっている。図１のような出力を行うためには、複数の技術文書から抽出された要素技術や効果の表記の揺れを同定する必要がある。例えば、自然言語処理分野で頻繁に要素技術として利用される「サポートベクトルマシン」は“SVM”や“Support Vector Machine”と表記されることがある。また、効果の表現に関して、「精度が向上」と「解析精度が改善」はほぼ同一内容であると思われるが、現状ではその同定処理を自動的に行うまでには至っていない。なお、このような処理を行う研究もすでに一部で始まっている [4]。

4 おわりに

本稿では、第７回および第８回 NTCIR ワークショップ特許マイニングタスクの概要を述べ、このタスクのテストコレクションを用いて筆者らが開発したシステムを紹介した。

[参考文献]

[1] 村田真樹，Saeger, S.D.，橋本力，風間淳一，山田一郎，黒田航，馬青，相澤彰子，鳥澤健太郎，論文データからの重要情報の抽出と可視化，第 23 回人工知能学会全国大会，2009.

[2] Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T., Overview of the Patent Mining Task at the NTCIR-7 Workshop, In Proceedings of the 7th NTCIR Workshop Meeting, pp.325-332, 2008.

[3] Nanba, H., Fujii, A., Iwayama, M., and Hashimoto, T., Overview of the Patent Mining Task at the NTCIR-8 Workshop, In Proceedings of the 8th NTCIR Workshop Meeting, pp.293-302, 2010.

[4] 西山莉紗，竹内広宜．同じ効果を持つ複数技術を同定するための知識抽出，第 24 回人工知能学会全国大会，2010.