NTCIR-7 Patent Mining Task 実施に向けて

2007.11.11
広島市立大学大学院 情報科学研究科
難波英嗣
E-mail: nanba@hiroshima-cu.ac.jp

1. 論文抄録分類タスクの提案

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てきた今日、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」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いる。この計画で、大学研究者の利用を想定した特許・論文情報統合検索 システムの整備が含まれていることからも、このような傾向が今後さらに強まっていくことがうかがわれる。

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、一般に無効資料調査と呼ばれている。同様の調査は、民間企業では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社内で行われることもある。そこで、NTCIR-7 特許マイニングタスクでは、特許と論文を対象にした検索や技術動向分析など、様々な目的に利用可能な言語処理技術の開発を目指す。本稿は、NTCIR-7 特許マイニングタスク実施のために行った予備実験について報告する。

なお、本稿は特許マイニングタスクを設計するにあたり、オーガナイザ間でディスカッションを行うために 2007 年 6 月 4 日に執筆されたものをベースにしているため、NTCIR-7 特許マイニングタスクの最終的な実験条件および評価方法は変わる可能性がある。また、実験に用いるシステムはあくまでも一例であり、タスク参加者はこれに縛られる必要はない。

2. 実験

本実験では、日本語論文抄録を特許分類体系のひとつである「国際特許分類」に自動分類することを目的とする。

2.1 システム

今回実験に用いるシステムは、以下の手順で入力された論文抄録の IPC を決める。

1. 入力された論文の抄録を形態素解析し、名詞、動詞、形容詞を抽出。
2. 手順 1 で抽出された形態素を検索クエリとし、汎用連想計算エンジン GETA (http://geta.ex.nii.ac.jp)を用いて、関連特許を検索し、検索結果上位 1000 件を得る。なお、この検索には著者が NTCIR-6 に参加した時のシステム[Nanba 2007]

を用いた。

3. 手順 2 で得られた 1000 件から、各特許に付与された IPC コードを抽出。
4. 手順 3 で得られた IPC コードについて、コード別に以下の式を用いてスコアを計算。

Score(IPC)=Σ(クエリに対する各特許の適合度)

5. スコアの高い順に IPC コードを出力。

手順 4 において、例えば、検索結果上位 1000 件中、"G01N 29/24"という IPC コードを含んだ特許が 10 件、250 件、570 件目に含まれており、それぞれの検索クエリに対する適合度が 0.7, 0.4, 0.2 であった時、Score(G01N 29/24)=0.7+0.4+0.2=1.3 となる。

システムの出力例を以下に示す。左から順に、トピック番号(NTCIR 論文データの ID)、ダミー、IPC、ランク、スコアを示す。また、参考までに、gakkai-0000266244 の正解は、図で下線が引かれている G01N 29/22、G01N 29/10 と G01B 17/00 の 3 件である。

gakkai-0000266244	1	G01N29/24	1	29.5213002788173
gakkai-0000266244	1	A61B8/00	2	28.1150048073693
gakkai-0000266244	1	G01N29/22	3	25.3221714568045
gakkai-0000266244	1	G01N29/26	4	20.748764030835
gakkai-0000266244	1	G01N29/10	5	20.6159443524759
gakkai-0000266244	1	H01J37/28	6	19.9997641742568
gakkai-0000266244	1	G01B7/34	7	17.2353879593635
gakkai-0000266244	1	H04R17/00	8	14.2459620492792
gakkai-0000266244	1	G01B21/30	9	12.3470176544705
gakkai-0000266244	1	A61B8/12	10	10.9183087409918

2.2 実験条件

- **テキストデータ：**

  公開特許広報 1993 年～2002 年

  NTCIR-1, 2 の抄録データ 46 万件

- **正解データ：**

  同一内容の特許と論文を人手で対応付けしたデータ 91 対を用いる。特許の分類番号を、対応する論文の特許分類番号と見なす。

  トピック数：91、正解 259(2.8 個/トピック)

  同一内容の特許と論文の対応付けは、以下の手順で行っている。

  1. 特許データベースから、「新規性喪失の例外の表示」の項目を含む特許を抽出(1993 年～2002 年のデータに約 9000 件)。
  2. この項目から、学会名と大会名を抽出。例えば、以下の例の場合、「情報処理学会」

と「第６０回全国大会」を抽出。

> 【新規性喪失の例外の表示】特許法第３０条第１項適用申請有り２０００年３月１４日　社団法人情報処理学会発行の「第６０回（平成１２年前期）全国大会講演論文集（４）」に発表

3. さらに、特許から発明者、出願年を抽出。
4. 手順 2、3 で抽出された情報を用い、NTCIR-1, 2 の抄録データから対応する論文の候補を抽出。(平均 10 抄録/トピック)
5. 手順 4 で抽出された候補を人手で判定。

論文に IPC を付与するのは専門的な知識を必要とするが、同一内容の特許と論文を対応付けるという作業であれば、専門知識を持っていない大学生でも容易に判定できる。同一内容の特許と論文データの対応付けデータは、2006 年度、難波の研究室の修士課程の学生が研究しており、正解データの作成に関してはある程度ノウハウは蓄積されている[末永 2007]。実際に、どのくらい対応がとれそうかについては正確には分かっていないが、上述のとおり、「新規性喪失の例外の表示」の項目を含む特許が 9000 件はある。このうち、少なくとも 1000 対程度は NTCIR-1, 2 のデータと対応付けできるのではないかと見込んでいる。

- **評価尺度：**

  これまでの特許分類タスクのようなカテゴリの階層を考慮した評価方法も考えられるが、今回はとりあえず以下の 2 つの尺度で評価を行った。また、計算には trec_eval Ver.8.1 を用いた。

  - MAP
  - 上位 n 件精度

2.3 結果

MAP 値は 0.1798、上位 n 件精度は表 1 のとおりである。

表 1　上位 n 件精度

上位 n 件	精度
5	0.1165
10	0.0780
15	0.0608
20	0.0544

2.4 考察

定量的には評価していないが、結果をざっと見る限りでは IPC のサブグループ(例えば G01N 29/24)までは完全には一致していないが、メイングループ(例えば G01N 29/)まで、あるいはサブクラス(例えば G01N)一致しているというケースがあった。このような場合に部分点を与えるという方法で評価するという方法も考えられる。

3. 想定される技術、アプローチ

- 論文用語、特許用語の相互変換技術[釜屋 2007]
- シソーラスの構築およびシソーラスベースのクエリ拡張[Nanba 2007]
- 一般的な文書分類技術
- 機械学習の適用

4. おわりに

NTCIR-7 Patent Mining Task 実施に向けて、論文への特許分類付与タスクを提案し、現実的にどの程度実現できそうか検証した。入力された論文抄録に対し、66,007 個の IPC コードからひとつ以上のコードを付与するという設定で実験を行った結果、MAP 値 0.l798 が得られた。

今回の実験に用いたシステムは、難波が NTCIR-6 検索タスクに参加した時のものをベースに 1～2 時間程度時間をかけて作成した。システムの構築には色々な方法があるとは思うが、これまで特許検索タスクに参加していたグループは、あまり労力をかけることなく新タスクに参加することができる。また、IPC ではなく F タームによる分類という設定にすれば、分類タスクに参加していたグループも比較的参加しやすいのではないかと思う。また、F ターム分類の場合、評価方法やツールなど、これまでのリソースが流用できると思われる。

参考文献

釜屋英昭，難波英嗣，奥村学，新森昭宏，谷川英和，鈴木泰山 (2007) “特許、論文間の引用情報を用いた論文用語の特許用語への変換” 情報処理学会自然言語処理研究会, NL-178，97-102.

Nanba, H. (2007) “Query Expansion using an Automatically Constructed Thesaurus” in Proceedings of the 6th NTCIR Workshop, 414-419.

末永健 (2007) “同一内容の特許と論文の対応付け” 広島市立大学情報科学研究科修士論文.