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

Technical Trend Analysis for Research Papers and Patent: Patent Mining Task of the Seventh and the Eighth NTCIR Workshops

難波英嗣 [1]，藤井敦 [2]，岩山真 [3]，橋本泰一 [4]

Hidetsugu NANBA[1]; Atsushi FUJII[2]; Makoto IWAYAMA[3]; Taiichi HASHIMOTO[4]

1 広島市立大学大学院 情報科学研究科（〒731-3194 広島市安佐南区大塚東 3-4-1) Tel: 082-830-1584 E-mail: nanba@hiroshima-cu.ac.jp

2 筑波大学大学院図書館情報メディア研究科（〒305-8550 茨城県つくば市春日 1-2) E-mail: fujii@slis.tsukuba.ac.jp

3 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（〒185-8601 東京都国分寺市東恋ヶ窪 1-280） E-mail: makoto.iwayama.nw@hitachi.com

4 東京工業大学統合研究院（〒226-8503 横浜市緑区長津田町 4259) E-mail: hashimoto@iri.titech.ac.jp

著者抄録

本稿では，第 7 回および第 8 回 NTCIR ワークショップにおいて実施された特許マイニングタスクと，このタスクで構築されたテストコレクション(評価用ベンチマーク)について述べる。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，技術動向マップを自動的に作成することである。本稿では，特許マイニングタスクで実施された 2 つのサブタスク：(1)学術論文分類と(2)技術動向マップ作成について説明する。また，国際的に利用されている特許分類体系のひとつである国際特許分類(IPC)に，学術論文を自動分類するシステムを紹介する。

Author Abstract

This paper introduces the Patent Mining Task of the Seventh and the Eighth NTCIR Workshops and the test collections produced in this task. The task's goal is the creation of technical trend maps from a set of research papers and patents in a particular research field. In this paper, we explain two subtasks: (1) research papers classification and (2) technical trend maps creation, which are conducted in the Patent Mining Task. We also show a system that classifies research papers into the International Patent Classification (IPC) system, which is a global standard hierarchical patent classification system.

キーワード
特許，学術論文，文書分類，情報抽出，NTCIR，評価ワークショップ
Key words
patent, research paper, text classification, information extraction, NTCIR, evaluation workshop

1. はじめに

本稿では，国立情報学研究所(NII)が主催する評価ワークショップ NTCIR において，筆者らが行っている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(評価用ベンチマーク)の構築研究について述べる。評価ワークショップとは，複数の研究グループが協調と競争を通して，問題設定やテストコレクション，評価方法について共同開発していく枠組みである。

筆者らは本ワークショップにおいて「特許マイニングタスク」を企画し，国内外から参加グループを募り，2007 年から研究を開始した。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，2009 年 6 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 2009」においても，推進計画 2006，2007 および 2008 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。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，このような傾向は今後さらに強まっていくと思われる。

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，本ワークショップの特許マイニングタスクである。

本タスクでは，日本語または英語論文抄録に，特許分類体系のひとつである「国際特許分類」(International Patent Classification: IPC)を自動的に付与し，さらに，同一分類の特許から技術動向マップを自動的に作成することを目的とする。特許を自動分類するタスクとしては，これまでに第 5 回 [1]および第 6 回 [2]NTCIR ワークショップにいて F ターム分類タスクが実施されてきたが，今回のタスクでは，分類対象となる文書が特許から論文に変わるため，特許と論文で使われる用語の違いについて新たに検討する必要がある。

特許では請求範囲をなるべく広く確保するため，一般性の高い特許用語を用いて記述する傾向がある。また，特許では学術用語よりも多様な表現が用いられることが多い。例えば，「機械翻訳」という学術用語に対する特許用語は「機械翻訳」の他にも「自動翻訳」「言語変換」などがある。このため，単純に表層的な単語の一致度を用いる従来の分類モデルでは，十分な分類精度が得られるとは限らない。本タスクでは，このような論文と特許の用語の使われ方の違いを吸収できる分類，検索，分析のための基礎技術の確立を目指している。

2. 関連研究

特許と論文を横断的に検索するための研究として Nanba ら[3]の研究が挙げられる。近年，特許中で関連論文を，また論文において関連特許を引用するケースが増えているが，このような文書間の引用関係をたどれば，論文や特許と関連する文書を集めることができる。そこで Nanba らは特許中で関連文献が引用される「従来の技術」という項目を解析して引用論文の書誌情報を抽出し，特許と論文間の引用関係を解析している。ただ，特許中の引用文献の中で論文が占める割合と，論文中の引用文献の中で特許が占める割合は数パーセント程度であるため，あるテーマに関する特許と論文を網羅的に収集するのに，引用関係をたどるだけでは十分とは言えない。

特許と論文を横断的に検索するための別のアプローチとして，難波ら[4]は，論文用語を特許用語に自動変換する手法を提案している。例えば，論文用語「フロッピーディスク」を特許用語「磁気記録媒体」に自動変換する。難波らは，論文用語の特許用語への変換を実現するため，特許と論文間の引用関係に着目している。一般に，引用関係にある特許と論文は，同一トピック(分野)である可能性が高い。そこで，ある用語を表題に含んだ論文を収集し，それらと直接引用関係にある特許から，特許のトピックを示す用語を抽出すれば，入力された論文用語に関連する特許用語の変換が実現できる。

特許と論文を対象としたこの他の研究として，TREC Chemistry Track[注1)]が挙げられる。これは，評価ワークショップの一つである TREC において 2009 年より新しく始まったタスクのひとつであり，化学分野の論文と特許に特化したジャンル横断検索を目的としている。その研究の詳細は，2009 年 11 月に開催される会議で報告される予定である。

3. 特許マイニングタスク

本章では，特許マイニングタスクの概要について述べる。3.1 節では特許マイニングタスクを実施している NTCIR ワークショップについて，3.2 節では特許マイニングタスクのロードマップについて，3.3 節ではサブタスクの概要について，3.4 節では評価について，それぞれ述べる。

3.1 NTCIR ワークショップ

本節では，テストコレクションの重要性と NTCIR ワークショップの意義について述べる。情報検索や自然言語処理などの言語情報処理に関する研究では「情報要求」「言葉の意味」「感情」といった，厳密な定義が困難な概念を研究の対象としている。しかし，科学や工学における一つの研究分野として言語情報処理を位置付けるためには，問題の定式化や評価において，学問として要求される水準を満たす必要がある。すなわち，学術研究としての実証性，客観性，再現性が求められている。そこで，複数の研究者が共有できる評価基盤としてのベンチマーク(テストコレクション)が重要性を増している。大規模で再利用可能なテストコレクションを組織的に構築するために，評価ワークショップという活動形態が存在する。評価ワークショップでは，複数の研究グループが互いに競い合いながら問題設定，テストコレクション，評価方法を開発する。

NTCIR は一年半の周期で開催される NII 主催の評価ワークショップである。ただし，研究発表だけの場ではない。オーガナイザから提供されたデータを用いて，参加者が共通の「研究課題(タスク)」を実行し，

注1) https://wiki.ir-facility.org/index.php/TREC_Chemistry_Track

互いのシステムを比較評価するための場である。筆者らは,第 7 回 NTCIR ワークショップ(NTCIR-7: 2007 年 6 月～2008 年 12 月)から，特許と論文を処理対象としたテストコレクションの構築研究として特許マイニングタスクを開始しており，現在，第 8 回ワークショップ(NTCIR-8: 2009 年 1 月～2010 年 6 月)において参加者を募集中である。次節では，NTCIR-7 および NTCIR-8 特許マイニングタスクについて述べる。

3.2 特許マイニングタスク ロードマップ

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 1 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ある。このような技術動向マップを自動生成するツールは，第 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 AAA] [特許 XXXXXX]		[論文 BBB]
要素技術 2	[論文 CCC]		
要素技術 3		[特開 YYYYYY]	[特許 ZZZZZZ] [特許 WWWWWW]

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

このうち,「学術論文分類サブタスク」については，すでに NTCIR-7 で実施している。また，NTCIR-8 ワークショップでは, 2 つのサブタスクを実施する。図 2 に特許マイニングタスクのロードマップを示す。次節では，各サブタスクの概要について述べる。

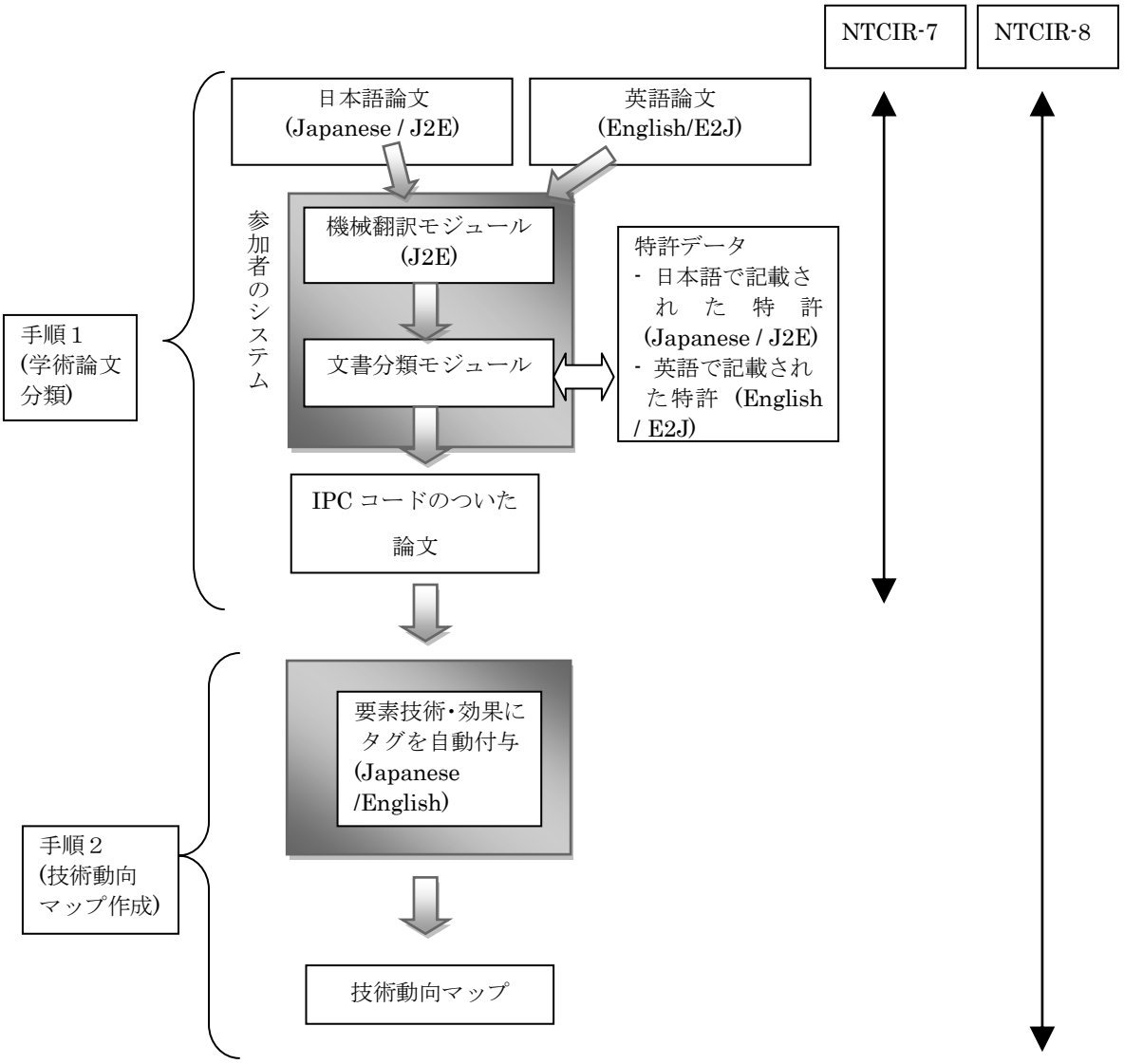

図 2 特許マイニングタスクロードマップ

3.3 サブタスクの概要

3.3.1 学術論文分類サブタスク

前述のとおり，特許マイニングタスクでは日本語または英語論文抄録に，特許分類体系のひとつである 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 [注2)]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 3 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

[注2)] NTCIR-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。

れぞれ示している。タスクの参加者は，図 3 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 3 システムの入力例

学術論文分類サブタスクでは，以下に示す 4 つの課題を実施する。

- 日本語文書分類(Japanese)：日本語の論文を日本語で記載された特許データを用いて分類する。
- 英語文書分類(English)：英語の論文を英語で記載された特許データを用いて分類する。
- 言語横断文書分類(J2E)：日本語の論文を英語で記載された特許データを用いて分類する。
- 言語横断文書分類(E2J)：英語の論文を日本語で記載された特許データを用いて分類する。

以上のサブタスクは図 4 にまとめられる。

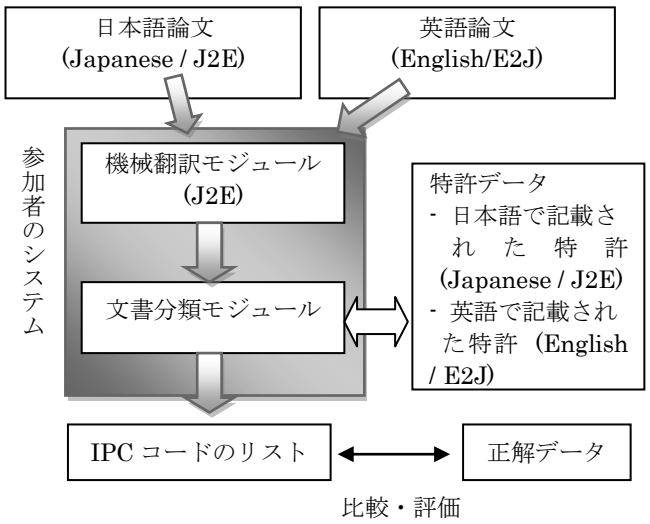

図 4 学術論文分類サブタスクのまとめ

3.3.2 技術動向マップ作成サブタスク

このサブタスクは，NTCIR-8 から開始する課題であり，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 5 に示すように，要素技術と効果を示す箇所に，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 5 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 5　技術動向マップ作成サブタスクに用いるデータの一例

技術動向マップ作成サブタスクでは，次の課題を実施する。

- 日本語情報抽出(Japanese)：日本語の論文と特許から，要素技術と課題に関する箇所を自動抽出する。
- 英語情報抽出(English)：英語の論文と特許から，要素技術と課題に関する箇所を自動抽出する。

NTCIR-8 特許マイニングタスクでは，これらの課題に加え，技術動向可視化サブタスクも実施する。このサブタスクは，技術動向マップ作成サブタスクの一種であるが，技術の効果について，Value タグが付与される表現が数値に限定されている点が，技術動向マップ作成サブタスクと異なる。図 6 は，技術動向可視化サブタスクに用いるデータの一例である。

<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 6　技術動向可視化サブタスクに用いるデータの一例

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 5 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 7 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 6 のようなグラフは十分に有用である。

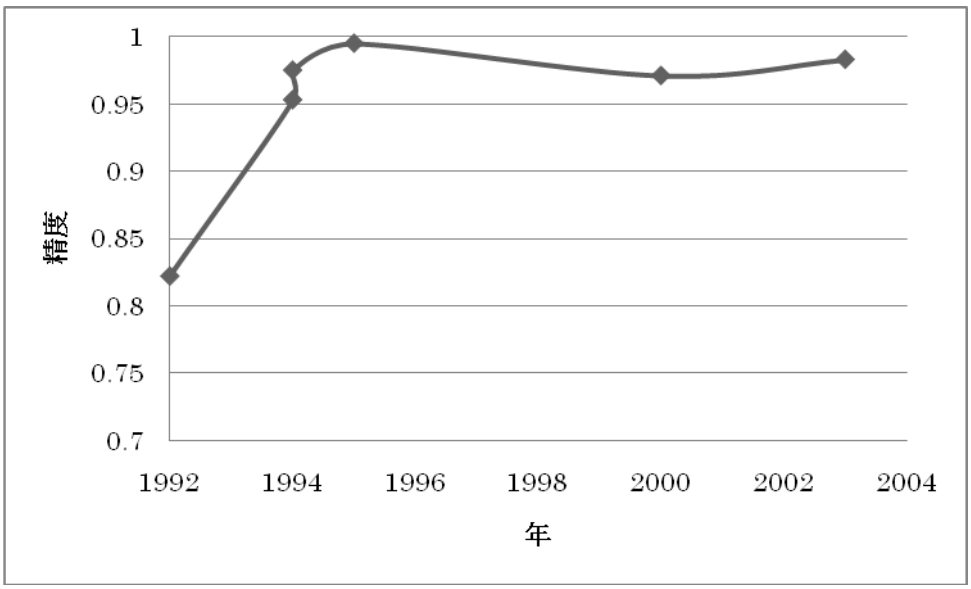

図 7 技術動向可視化サブタスクの出力例

3.4 評価

本節では，すでに NTCIR-7 特許マイニングタスクで実施した学術論文分類サブタスクの評価について述べる。

- **文書データ**

NTCIR-7 では，タスク参加者には，表 1 の文書セットが事前に配布された。また，この他に，データ(1)～(3)の各文書に人手で IPC コードを付与したデータが訓練用データとして配布された。

表 1 文書セット

データ名	年	サイズ	文書数	言語
(1) 日本国公開特許公報	1993–2002	100 GB	3.5M	日
(2) 米国特許	1993–2000	33 GB	0.99M	英
(3) Patent Abstracts of Japan (PAJ)	1993–2002	4.2 GB	3.5M	英
(4) NTCIR-1，NTCIR-2 言語横断タスクテストコレクション(論文抄録データ)	1988–1999	1.4 GB	0.26M	日/英

- **正解データ**

976 論文に人手で IPC を付与したデータを用意し，評価に用いた。なお，正解データの作成に関する詳細は，文献 [5]を参照されたい。

- **参加グループ**

日本語文書分類には 24 システム，英語文書分類には 20 システム，言語横断サブタスク(J2E)には 5 システムからの結果が，それぞれ提出された。また，参加グループ数は計 12 グループであり，その内訳は表 2 に示すとおりである。

表 2 参加グループ内訳

	日本	アジア（日本以外）	欧州	北米
大学	3	4	0	2
企業	2	0	1	0

なお，評価方法や評価結果の詳細は，ここでは割愛する。代わりに，次章で日本語文書分類への参加システムを紹介する。評価結果の詳細は，筆者らの報告書[5]を参照のこと。

4. 学術論文自動分類システム動作例

本章では，NTCIR-7 特許マイニングタスクの成果の一例として，筆者らが構築した学術論文自動分類システム[6]を紹介する。4.1 節ではシステムの動作例を，4.2 節では自動分類の仕組みについて，それぞれ説明する。

4.1 学術論文自動分類システム

図 8 は，学術論文自動分類システムの入力画面である。この画面内の検索フォームに論文の概要を入力し，「検索」ボタンを押すと，その概要に関連する IPC コードが自動的に列挙される(図 9)。ここで，図 9 中の「スコア」は，4.2 節で述べる方法によりシステムが計算した論文概要と各 IPC コードとの関連度を示す。また，図中の「用語リスト」をユーザがクリックすると，各 IPC コードの特徴を示す代表的な専門用語が一覧表示される。

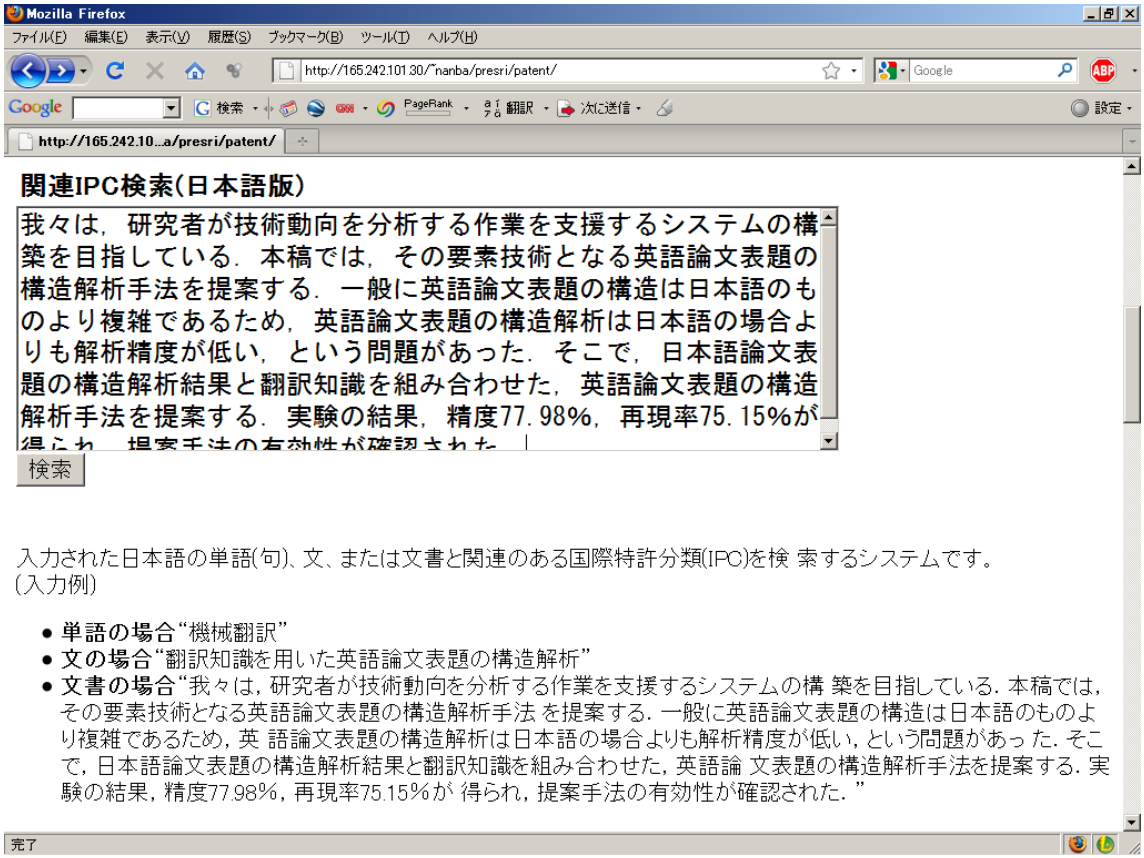

図 8 学術論文分類システム動作例(入力画面)

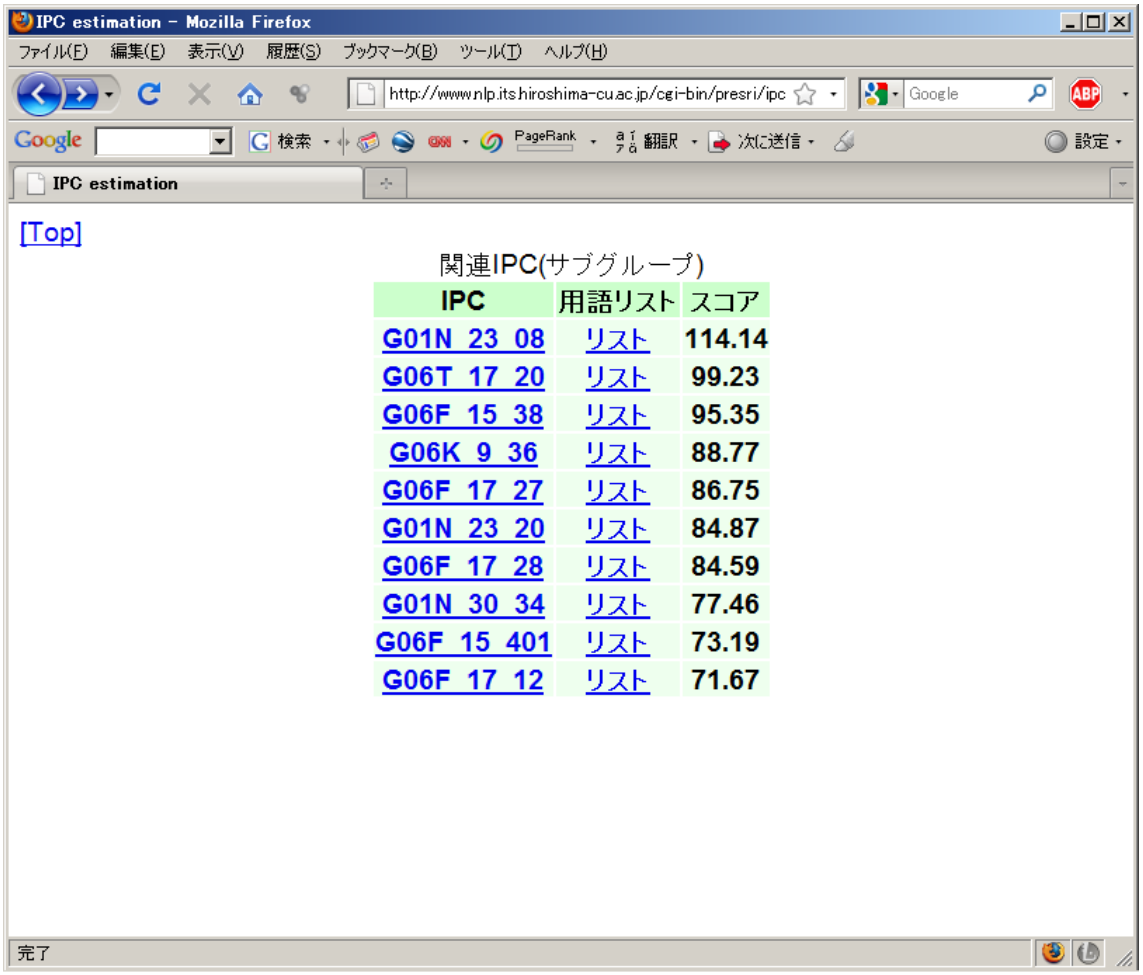

図 9 学術論文分類システム動作例(実行結果)

4.2 自動分類システムの仕組み

この分類システムでは，k-Nearest Neighbor 法(k-NN 法)という手法を採用している。k-NN 法を用いた学術論文の国際特許分類への自動分類手順について，図 10 を用いて説明する。本システムでは，学術論文が入力されると，まず，論文から内容語(名詞，動詞，形容詞)を自動抽出する。次に，それらの内容語を検索キーワードとして，特許検索システムを用いて関連特許を検索する。一般に，検索結果として得られた特許集合にある IPC コードが数多く付与されていれば，その IPC コードは入力された論文と関連度が高いと考えられる。そこで，検索結果上位 170 件[注3)]の各特許に付与された IPC コードを抽出し，コード別に以下の式を用いてスコアを計算し，スコアの高い順に IPC コードを出力することにより，図 8 に示す結果が得られる。

$$\text{スコア(IPC)} = \Sigma(\text{学術論文に対する各特許の類似度})$$

注3) この値は、予備実験により決定した。

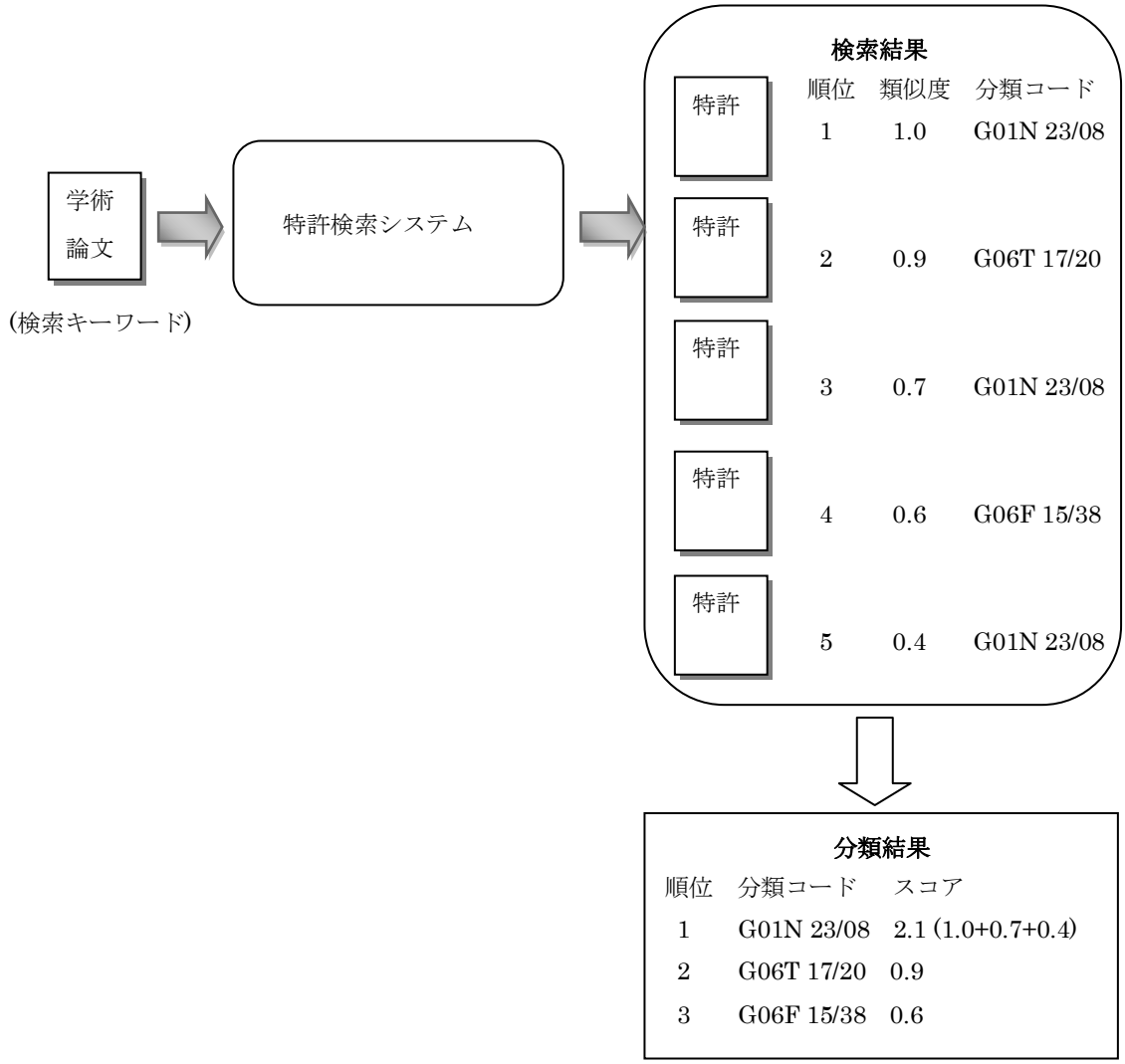

図 10 k-Nearest Neighbor 法を用いた学術論文分類システム

このシステムが，実際にどの程度実用に耐えうる物かを調べるため，分類システムが出力した上位 n 件の分類コードが, 人間の付与した IPC コードをどの程度正しく抽出できているか(再現率)により調べた。結果を表 3 に示す。この結果より，上位 1 件で約 20%，上位 10 件で約 60%の IPC コードが正しく付与できていることがわかる。特許と論文を対象にした技術動向分析の支援を行うためには，上位 1 位における再現率のさらなる向上が必要であるものの，今回の結果は，特許の検索初心者にとってはある程度有効であると考えられる。一般に，特許を効率的に検索するためには，検索キーワードに加え IPC などの特許分類コードも併用される。しかし，検索初心者にとって，適切な特許分類コードの選択そのものが困難であり，これにはある程度の技術と経験が必要とされる。このような場合，ユーザが本システムに調べたい分野の論文を入力すれば，その論文と関連する IPC コードが列挙される。表 3 から，ユーザが結果の上位 10 件まで見れば，60%以上の確率で該当する IPC コードが得られることから，特許検索初心者に対する IPC コードを用いた特許検索の敷居をある程度下げる効果があり，検索支援につながると思われる。

表 3 上位 n 件の再現率

順位	再現率
1	19.4(216)
2	31.0(346)
3	38.3(428)
4	43.7(487)
5	47.6(531)
10	58.6(653)
20	69.8(778)
50	80.9(902)
100	84.3(940)
500	84.7(944)
1000	84.7(944)

5. おわりに

本稿では，NTCIR-7 および 8 において，筆者らが実施している「特許マイニングタスク」について報告した。NTCIR-7 で構築した評価用テストコレクションは一般公開されており，国立情報学研究所と覚書を交わせば入手できる。詳しくは，NTCIR の Web ページを参照されたい[注4)]。また，NTCIR-8 特許マイニングタスクは，2009 年 5 月から参加募集を開始している。このタスクに興味のある読者は，特許マイニングタスク Web ページ[注5)]を参照のこと。参加受付も随時行っているので，希望者は特許マイニングタスクオーガナイザ(ntcadm-mining@slis.tsukuba.ac.jp)まで連絡をいただきたい。

参考文献

1) Iwayama, Makoto; Fujii, Atsushi; Kando, Noriko..”Overview of Classification Subtask at NTCIR-5 Patent Retrieval Task”. Proceedings of the 5th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access. Tokyo, 2005-12-06/09, National Institute of Informatics. 2005, p. 278-286.

2) Iwayama, Makoto; Fujii, Atsushi; Kando, Noriko. “Overview of Classification Subtask at NTCIR-6 Patent Retrieval Task”. Proceedings of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access. Tokyo, 2007-05-15/18, National Institute of Informatics. 2007, p. 366-372.

3) Nanba, Hidetsugu; Anzen, Natsumi; Okumura, Manabu. Automatic Extraction of Citation Information in Japanese Patent Applications. International Journal on Digital Libraries. 2008, vol. 9, no. 2, p. 151-161.

注4) http://ntcir.nii.ac.jp

注5) http://www.ls.info.hiroshima-cu.ac.jp/~nanba/ntcir-8/cfp.html

4) 難波英嗣, 釜屋英昭, 竹澤寿幸, 奥村学, 谷川英和, 新森昭宏. 論文用語の特許用語への自動変換. 情報処理学会論文誌データベース. 2009, vol. 2, no.1, p. 81-92.

5) Nanba, Hidetsugu; Fujii, Atsushi; Iwayama, Makoto; Hashimoto, Taiichi. “Overview of the Patent Mining Task at the NTCIR-7 Workshop”. Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access. Tokyo, 2008-12-16/19, National Institute of Informatics. 2008, p. 325-332.

6) Nanba, Hidetsugu. “Hiroshima City University at NTCIR-7 Patent Mining Task”. Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-lingual Information Access. Tokyo, 2008-12-16/19, National Institute of Informatics, 2008, p. 369-372.