F-031

階層分類による科学技術論文抄録と国際特許分類関連性決定手法の提案

Hierarchical Classification Approach for Detecting Relevance between Scientific Abstract and International Patent Classification (IPC)

チャオヤン†　青野雅樹‡
Chao Yang　Masaki Aono

1.　はじめに

特許は専門用語が多く文章も多様で文献数も多いなど，特許文献は特殊なため，特許に関連する作業は大きな人的コストがかかる．それを改善するため，特許に関する研究は過去に多く行われてきた．特許の分類に関しては NTCIR というワークショップなどで広く研究を行われている．

本報告では研究者が新発想や新発見により新たなものや技術を開発したときに，その科学技術論文抄録と国際特許分類の関連性を検出することを提案する．本提案手法は，はじめに，NTCIR-7 英語版の特許データを用いて，階層カテゴリを作成する，作成した階層カテゴリより，対象の論文抄録と国際特許分類(IPC*)の関連を検出する．検出した IPC をランキングして上位の 1000 件取り出し，評価に関する研究とその実験結果を報告する．

2.　関連研究

論文抄録から特許への分類は NTCIR ワークショップで多く研究されている。Hanif [1]らはオントロジー技術を用いて分類する，Tong Xiao[2] らは KNN とランキング学習により分類するなどがあげられる．本研究では世界知的所有機関(WIPO**)を提供している IPC 階層カテゴリを用いて，階層カテゴリを作成し，与えられた論文抄録を階層カテゴリに分類行う手法を提案する．特許のセッションからサブクラスまで関連した特許群へ分類することを目的としている．

3.　国際特許分類(IPC)

国際特許分類の分類記号はセクションからサブグループに至る階層を示すアルファベットと数字の組み合わせである．約 8 万カテゴリが階層的に構成されている．以下の表 1 は国際特許コードの階層の例を示す．

表 1　国際特許分類

A	01	B	1	/02	
セックション					8セックション
クラス					128クラス
サーブクラス					648サブクラス
メイングループ					7200メイングループ
サーブグループ					72000サブグループ

4.　提案手法

4.1　提案システム構成

これまでのテキスト分類や NTCIR ワークショップが行われていた特許分類の研究はフラット分類に関する研究が多く行われていた．しかし，フラット分類ではカテゴリが多くなった場合分類コストがかかる．本研究では高速分類することを考慮し，階層構造カテゴリを用いて，階層分類手法を提案する．提案システム構成は以下の図 1 のようにシステムを提案する．

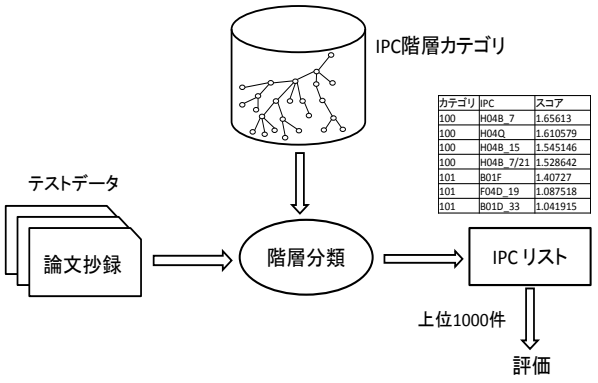

図 1　システムの構成

4.2　階層カテゴリの作成

世界知的所有機関[3](WIPO)を提供している IPC 階層構造に従って，階層カテゴリを作成する．

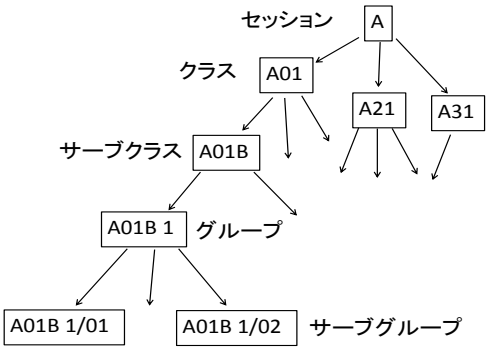

図 2　階層カテゴリ

階層カテゴリの素性に関しては NTCIR-7 英語版のデータを用いて，それぞれのカテゴリに関連した特許に含まれる単語を階層ごとに素性とする．本研究では一部の約 2 万 1 千件の特許を用いて，階層カテゴリを作成し，検証実験を行う．

4.3　階層分類におけるノードのスコア付け

階層分類の入力となる対象論文が国際特許分類のどのカテゴリに関連するか階層分類を用いて行う．まず，ルートノードカテゴリから始まり，各ノードカテゴリでの素性と与えられた論文抄録の素性もっとも共起素性が表れるカテ

†豊橋技術科学大学　大学院　情報・知能工学専攻
‡豊橋技術科学大学　情報・知能工学系

**WIPO: World International Patent Organization
*IPC: International Patent Classification

ゴリを次のサブノードカテゴリとして選択する．同時に関連度のスコアも計算する．このような手順で論文抄録から国際特許分類へ階層分類を行う．また，関連度のスコア(RS)は共起素性の素性数やノードの素性数を用いて式（1）で定義する．

$$RS(d, node) = \frac{n(q \cap node) \times \max(tf(q \cap node))}{n(q \cup node)} \quad ....(1)$$

ここで，q は分類対象論文の素性，$node$ はノードの素性．$n(q \cap node)$ は分類対象論文とノードでの共起素性の数．$\max(tf(q \cap node))$ は分類対象論文とノードの素性出現頻度である．$n(q \cup node)$ は分類対象論文とノードの集合素性数．

4.4 IPC リスト

国際特許分類（IPC）は細かく分けられており，1 つの論文抄録は多く国際特許分類(IPC)と関連していると考えられる。本研究ではスコアの高い上位 1000 件リストとして取り出して，評価する．

5. 検証実験

5.1 実験概要

本研究では論文と国際特許分類の関連を調査するために，NTCIR-7 を提供している dry-run データセット 97 カテゴリをテストデータとする．更に NTCIR-7 の特許データ 21000 カテゴリを抽出し，学習データとする．まず，論文抄録を形態素解析し，Stopword を削除し， POSTagger により固有名詞のみ取り出す．取り出した素性を Wordnet 辞書から上下関係概念を取り出して，その論文の素性とする．特許データは抄録の部分のみ抽出し，Stopword を削除し，POSTagger による形態素解析を行い，固有名詞のみ取り出し，素性とする．本実験ではサブクラス階層まで実験評価を報告する．

5.2 比較実験

提案手法と比較するため，情報分類や情報検索で使われている $tdidf$ 手法を用い，コサイン類似度により，論文と特許の類似度を計算し，上位 1000 件を取り出して評価を行う．本実験ではサブクラス階層まで分類する．

$$tfidf(t,d) = tf(t,d) \times \log(\frac{N}{df(t)})$$

ここで，$tdidf(t,d)$ 単語 t が文書 d に出現する重みベクトル，$tf(t,d)$ 単語 t が文書 d に出現する頻度である．$df(t)$ 全体文書において，単語(t)が出現する頻度，N 全体文書総を表している．

コサイン類似度は以下の式で用いて算出する．

$$\cos(d_1, d_2) = \frac{\vec{d}_1 \cdot \vec{d}_2}{|\vec{d}_1| \cdot |\vec{d}_2|}$$

ここで d_1，d_2 それぞれは文書 1 及び文書 2 のベクトルである．

本研究の分類精度を Recall-Precision の平均による評価を行った．

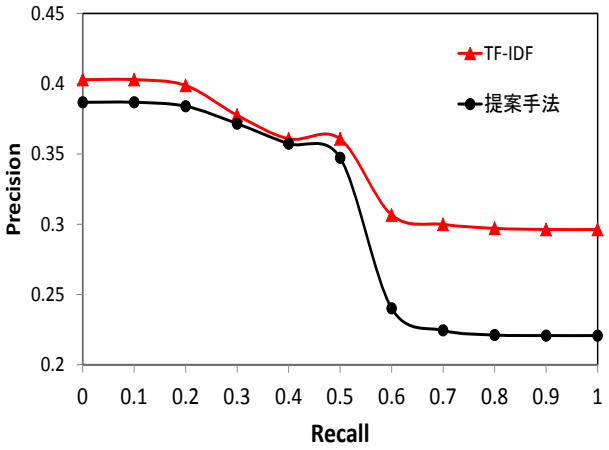

図 3 Recall-Precision グラフによる分類精度比較

提案手法では，従来フラット分類手法より悪化とみられる．これは，$tdidf$ 手法を用いてフラットを分類精度は全体文書数及び全素性を用いたため，結果が良くなると考える．本提案手法では，各サブカテゴリにおいての素性の考慮したため，精度が従来手法より低くなったと考える．

6. まとめと今後の課題

本研究では階層分類手法を用いて，科学技術論文抄録から国際特許への階層分類を行った．階層素性カテゴリにおいて，固有名詞のみ特定したため，素性数が少ないと考えられる．階層構造において，分類精度向上させるためには，各サブノードカテゴリにおいて，上下関係概念だけではなくて，Wordnet から類義語及び意味関係概念も抽出して素性とすることを考えられる．

本研究では国際特許分類のサブクラス階層しか分類を行ってなかったため，今後の課題はサブグループまで階層的に分類を行う予定している．また，最新の国際特許分類に適用できるため，NTCIR-10 データセットを用いて，検証実験を行うがあげられる．

7. 参考文献

[1] Md Hanif Seddiqui ,Yohei Seki, Aono Masaki. Ontology bas Approach to Patent Mining for Relating International Patent Classification (IPC) to a Scientific Abstract, Proceedings of NTCIR-7 Workshop Meeting, December 16–19, 2008, Tokyo, Japan

[2] Tong Xiao, Feifei Cao, Tianning Li KNN and Re-ranking Models for English Patent Mining at NTCIR-7, Proceedings of NTCIR-7 Workshop Meeting, December 16–19, 2008, Tokyo, Japan

[3] 世界知的所有機関（WIPO）
http://www.wipo.int/portal/index.html.en