情報処理学会論文誌 Vol.54 No.1 414–422 (Jan. 2013)

技術論文の国際特許分類体系への自動分類システムにおける機能要素の高度化と最適化

鈴木 克典[1,†1] 湯川 高志[1,a)]

受付日 2012年5月4日, 採録日 2012年10月10日

概要：企業が技術戦略を策定する際，技術動向を調査・視覚化してその方針決定に役立てる．このような調査に有益なツールの 1 つとして特許マップがあり，多くの場合，この作成には特許文書に付与されている国際特許分類（IPC）コードが利用される．この特許マップに学術論文の情報を追加することができれば，より有益なツールになると考えられるが，現在学術論文は IPC に基づいた分類が成されておらず特定分野の網羅的な調査には大きな負担がともなう．本稿では，学術論文に自動的に IPC 分類コードを付与するシステムの精度向上を目指し，分類手法を構成する 3 つの要素について高度化とパラメータの最適化を図った．本稿が最適化した 3 つの要素とは，文書からのキーワード抽出，クエリ拡張，類似特許検索における適合度計算である．これらの要素について，いくつかの方法とパラメータを変化させて実験的に評価し，その最適値を見出した．結果，再現率をほとんど低下させることなく，MAP 値を 0.199 から 0.494 に改善することができた．

キーワード：特許情報処理，特許マップ，文書分類，概念ベース

Improvement and Optimization of Automatic Research Paper Classification into an International Patent Classification System

Katsunori Suzuki[1,†1] Takashi Yukawa[1,a)]

Received: May 4, 2012, Accepted: October 10, 2012

Abstract: A patent map is a technical information map which was made by analyzed patents. It would become more useful if it includes information not only on patents but also research papers. To achieve this, automatic classification of research papers into the International Patent Classification System (IPC) is required, and some systems has been developed. In the present paper, the elemental functions which is consisted in an automatic technical paper classification system are improved and optimized for obtaining higher classification accuracy. The new method that focuses on keyword extraction is proposed and search query is expanded by concept base. In addition, patent document has a specific structure consisting sections represented in different granularity keywords. A method that weighing the score of similarity calculated between each structural element is proposed. Using the methods described above, the optimum value of precision is got by varying parameters. As a result, the MAP value was up to 0.494 from 0.199 in empirical assessment.

Keywords: patent information processing, patent map, document classification, concept base

1. はじめに

企業は技術戦略を策定する際，対象とする技術の研究開発動向を調査・視覚化し方針決定に役立てる．このような調査に有益なツールの 1 つとして特許マップがあり，これは特許文書を分析・整理して技術情報を二次元平面上に

[1] 長岡技術科学大学
Nagaoka Univercity of Technology, Nagaoka, Niigata 940–2188, Japan
[†1] 現在，NEC ソフト株式会社
Presently with NEC Soft, Ltd.
[a)] yukawa@vos.nagaokaut.ac.jp

マッピングしたものである [1]．この特許マップを作成するにあたっては目的の分野に沿った特許文書を収集する必要があり，そのために特許分類コードに基づいた検索が行われる．これには，あらかじめ特許文書に付与された国際特許分類（International Patent Classification，以下 IPC）等の体系化された分類コードが利用される．

ところで，技術の公表は，特許だけではなく学術論文をはじめとした技術文書によっても行われる．特許文書から作成された特許マップに，他の技術文書の情報を追加することができれば，より広範な調査が可能であると考えられる．特許文書以外の技術文書には共通した分類コードが存在しないため，特許文書以外の技術文書の調査ではキーワードを指定して関連する文書を検索する方法が一般的である．しかし，このような検索により得られる技術文書は，特許マップ作成のための特定分野の網羅的な調査のためには十分とはいえない．なぜなら，キーワード検索による調査は，そのキーワードを含む技術文書が分類に関係なく検索されるため，無関係な文書が多くなり，また，一方で表記ゆれ等による検索漏れも多いからである．

そこで，IPC 体系に基づく分類コード（以下 IPC コード）が技術文書へ事前に付与されていれば，技術文書も同様に特定分野の網羅的な調査が可能となり，さらに特許マップに技術文書の情報を追加することが容易となる．このようなタスクは国立情報学研究所が主催するワークショップ NTCIR において学術論文分類タスク（パテントマイニングタスク）として設定されていた [2]．

本稿では，まず，既存の手法を用いた基本システムを構築し，前述の NTCIR のタスクで提供されているテストコレクションを利用してその特性を評価する．評価結果に基づき，再現率は高いものの精度（MAP 値）が低いことを明らかにする．そこで，このシステムの機能要素である，キーワード抽出，クエリ拡張，類似特許検索における関連度の計算，の 3 つについて，より高度な手法を導入するとともに，それらに用いられるパラメータの最適化を図る．

本稿の構成は以下のとおりである．2 章では先行研究について述べる．3 章では基本システムの構成および性能評価について述べる．4 章ではシステムの精度を向上させるための各要素の高度化手法について述べ，5 章で各手法の評価とパラメータを最適値について報告する．最後に，6 章で本稿をまとめる．

2. 先行研究

パテントマイニングタスクは，NTCIR ワークショップの第 7 回（2008 年）と第 8 回（2010 年）の 2 回にわたって実施された [2], [3]．これらのワークショップに参加したチームのアプローチは，類似文書検索技術を用いた手法と，機械学習を用いた手法との 2 つに大別される．前者は，課題の学術論文と類似した文書を特許文書セットから検索し，検索された特許文書ごとにベクトル空間モデル等で計算された RSV（Retrieval Status Value）と IPC コードとを集計し，それらから K-NN（K-Nearest Neighbor）法 [4] を用いて学術論文へ付与する IPC コードを決定するものである．後者は，特許文書セットからキーワードを抽出し，それらの統計情報を特徴量として機械学習手法によりモデルを構築し，課題の学術論文に付与する IPC コードを決定するものである．

近年，自然言語処理における分類問題には SVM（Support Vector Machine）をはじめとした機械学習手法がよく用いられるようになってきた．しかし本タスクの場合，分類項目数が 7 万と非常に多く，多クラスに対応した SVM を用いてもその学習に多大なコストを要する．さらに，あらかじめ IPC コードが付与されているのは特許文書のみであるため訓練データとして特許文書を用いることになる．ところが，課題は学術論文であり記述形態が異なるため，学習の際のドメインと課題のドメインが異なることになる．このため，NTCIR ワークショップにおいては機械学習のみを用いたシステムはあまり精度が高くない傾向にあった．

第 7 回のワークショップにおいては，類似文書検索を利用した手法が主流であった．特に Mase らのグループは，検索クエリを拡張することで最も高い精度を示した [5]．Mase らは，提供される学術論文が表題および概要のみであることから，そこからクエリとして使用するキーワードを抽出するには文書長が不十分であると主張した．そこで，事前に課題文書と似ている文書を学術論文データベースから検索して，これに含まれるキーワードを検索クエリに加える手法を提案した．また，類似文書の検索にはベクトル空間モデルを用いている．

第 8 回のワークショップにおいては機械学習を利用したアプローチが増えた．なかでも Mase らのシステムは，MAP 値で 2 位のチームと約 9 パーセントポイントの差をつける等突出した性能であった [6]．このシステムは，類似特許文書をベクトル空間モデルによる類似検索で獲得することは以前と同様であるが，IPC コードを決定する際に新たにランク学習（learning to rank）を取り入れた．ランク学習とは，検索結果が正解データに近くなるように検索キーワードに対する重みを機械学習により決定し，この重みを利用して文書の最適なランクを推定することで，分類性能を改善させる手法である．また，Gang らのチームは特許文書の構造を文書ランキングに応用した．これは特定の構造に出現したキーワードに定数倍の重みを付加し RSV を計算する手法である [7]．

このように，各ワークショップで最も優秀な性能を示したシステムは類似文書検索をベースとしたアプローチを採用しており，依然として類似特許文書検索の高精度化は重要である．そこで本稿では，特許文書検索技術に着目し，さらなる改善を加えることにより精度の高い分類システム

の実現を目指す．

3. 基本システムとその性能評価

類似文書検索に基づく基本的な IPC コード付与手法の特性を調べるために，基本手法を用いたシステムを構築し評価した．本章では，そのシステムの構成と動作について述べ，評価手法，評価結果について報告する．

3.1 基本システムの構成および動作

本システムは，特許文書にあらかじめ付与されている IPC コードを利用し，課題として与えられる学術論文に対して 1 つ以上の IPC コードを付与する．システム構成を**図 1** に示す．

システムは，まず特許文書集合から入力された学術論文との RSV が高い文書 k 件を検索する．RSV は，文書集合からはキーワードとして名詞を抽出し，確率モデルの 1 つである Okapi BM25 [8] に基づき計算する．入力された学術文書を Q，検索された特許文書を D としたとき，RSV は以下の式で計算される．

$$RSV(D,Q)=\sum_{t\in Q}\left(f_{qt}\frac{3.0f_{dt}}{(0.5+1.5l_d/l_{avg})+f_{dt}}w_t\right) \quad (1)$$

ここで，f_{qt} は Q におけるキーワード t の出現頻度，f_{dt} は D におけるキーワードの出現頻度，l_d は D のキーワードの延べ数，l_{avg} は文書セットにわたる l_d の平均値である．

また，w_t は以下の式で計算される．

$$w_t=\log\frac{N-n_t+0.5}{n_t+0.5} \quad (2)$$

ここで，N は文書セットの文書総数，n_t はキーワード t を含む文書の数である．

次に，検索された k 件の特許文書の RSV および付与されている IPC コードを基に各 IPC コードのスコアを計算し，最終的に IPC コードをランク付けしその上位 1,000 件を出力する．IPC コード C に対するスコア s_C は以下の式で計算される．

$$s_C=\sum_{i=1}^{k}\delta(ipc(D_i)=C)RSV(D_i,Q) \quad (3)$$

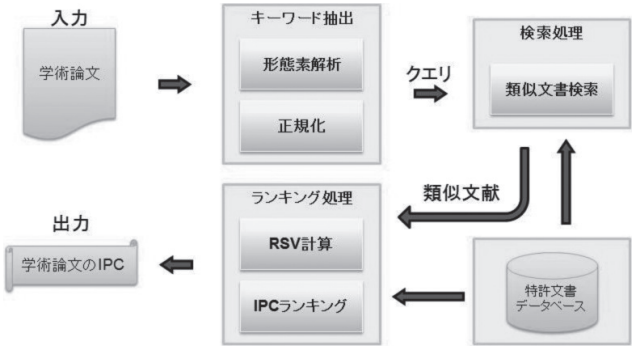

図 1 ベースラインシステム構成
Fig. 1 Baseline system architecture.

ただし，D_i は検索された k 件のうちの i 番目の特許文書である．また，$\delta(ipc(D_i)=C)$ は，特許文書 D_i に付与された IPC コードに C が含まれている場合には 1，そうでなければ 0 をとる関数である．

3.2 評価データおよび評価手法

本稿ではシステムの評価に，NTCIR から提供されるテストコレクションを利用している．このテストコレクションは，約 350 万件の公開特許公報（広報全文），200 件の学術論文（表題，および要約），および各学術論文に対する 1 つ以上の適合 IPC コードから構成されている．

評価指標についても NTCIR のパテントマイニングタスクに準じている．NTCIR で採用している評価指標は MAP 値である．MAP 値は，個々の課題文書に対する AP 値を求め，それをすべての課題文書にわたって平均した値として計算される．ある課題の学術文書 Q に対する AP 値は以下の式で計算される．

$$AP(Q)=\frac{1}{R_Q}\sum_{i=1}^{N}\frac{z_Q(i)}{i}\left(1+\sum_{j=1}^{i-1}z_Q(j)\right) \quad (4)$$

ここで，R_Q はその課題における正解の IPC コードの数，N はシステムが出力した IPC コードの数を表している．また，$z_Q(i)$ は，システムが出力した i 番目の IPC コードが正解に含まれていれば 1 を，そうでなければ 0 をとる．システムは，課題として与えられた学術論文に適合すると考えられる IPC コードを順位付けして，最大 1,000 件出力する．すなわち，本稿においては $N=1000$ である．

3.3 基本システムの評価結果

上述した基本システムを実装し評価した．その結果を**表 1** に示す．MAP 値に加え再現率（recall）についても評価結果を示している．

表 1 は IPC コードのレベル区分ごとの MAP 値および再現率を表している．IPC コードは Section から SubGroup の順で詳細な分類となる階層構造を有している．以降，単に MAP 値/再現率と記載した場合には，SubGroup レベルにおける各評価値を指すものとする．

学術論文に対する適合 IPC コード数は平均 2.4 件，最大でも 7 件であるのに対し，MAP の計算対象となるシステム出力の件数は最大 1,000 件である．その結果，表 1 から分かるように，再現率はどの IPC コードのレベルにおいてもほぼ 1 である．むしろ，このような条件において再現率

表 1 基本システムの性能
Table 1 Performance of baseline system.

Level	Section	Class	SubClass	Group	SubGroup
MAP	0.625	0.391	0.327	0.248	0.178
Recall	1.000	1.000	1.000	0.985	0.972

が低いのであれば，それは健全なシステムとはいえない．一方，MAP 値は，8 つのカテゴリしか設定されていない Section においても 0.625 と高くはないうえ，IPC コードのレベルが詳細になるにつれ急激に低下している．

本稿では再現率を高水準に維持したまま MAP 値を向上させるような改良を試みる．

4. 要素の高度化と最適化

本章では，基本システムのうち精度に関係すると考えられる 3 つの要素を高度化し，それらのパラメータの最適化を図る．まず，類似特許文書検索のためのキーワード抽出を高度化する．次に，類似特許の検索漏れを削減するためのクエリ拡張に概念ベースを利用する．最後に，RSV の計算において，特許文書が持つ固有の構造を利用する．

4.1 キーワード抽出

類似特許文書を検索するうえでキーワードの抽出手法は重要である．基本システムでは，他の一般的な検索システムと同様に，文書からキーワードを抽出するために形態素解析を利用している．ところで，特許文書では慣習的にカタカナ語の使用を避けて漢字による表記を用いることが多いため，一般的な文書と比較して長い漢字熟語が頻出するという特徴がある．こうした漢字熟語は，個々の形態素を組み合わせることで，それぞれの形態素の意味の集合を超えた新たな意味を獲得しており，それら個々のは形態素よりさらに強くその文書を特徴付けていると考えられる．

基本システムでは，キーワードを抽出する際に，漢字熟語を形態素の集合に分解してしまうため，このような，組合せとして持つ熟語の意味が失なわれてしまっている．漢字熟語が頻出する特許文書に対しては，個々の形態素よりは，むしろ漢字熟語，すなわち連続した名詞をすべて接続した結合名詞をキーワードとする方が自然である．

ところが，結合名詞をキーワードとした場合，それは文書を強く特徴付ける反面，検索漏れを引き起こすという問題がある．このため，語を結合してキーワードとして用いると，検索性能が低下する場合が多いことが報告されている [9], [10]．そこで，名詞をすべて結合させるのではなく，連続する N 個の名詞を結合したもの，すなわち単語 N-gram をキーワードとする手法を考える．このような手法もすでに提案されているが [11]，単語 N-gram のみをキーワードとしても良い精度は得られず，単語そのもの（すなわち unigram）もキーワードとして追加することで，精度の向上が得られると報告されている．しかしながら，上述のとおり，本稿が対象とする文書には複合語を多用した文章が多く，名詞が連続している場合にはそれらを結合した語のみをキーワードとして用いても分類性能の向上が得られる可能性が高い．そこで，名詞が連続している場合に，単語 N-gram のみをキーワードとして用いるものとする．

表 2 除去される接頭辞・接尾辞
Table 2 Prefixes and suffixes to be removed.

接頭辞	前記	当該	該				
接尾辞	等	側	法	群	形	部	間
	状	的	端	用	時	系	下
	係	内	面	性	外	付	中
	後	ごと	機	程度	上	毎	図
	付き	あたり	どうし				

また，結合名詞には，その前後に不要な接頭辞や接尾辞が共起しており，これも分類精度を低下させる原因になりうる．そこで事前に接頭辞や接尾辞を集計し，不要と判断される接頭/接尾辞は除去することとする．具体的には，大量の特許文書について単語の出現頻度を集計し，頻度の高い語について除去すべき接頭辞・接尾辞か否かを目視により判断して辞書を作成する．キーワード抽出において，これに照らし合わせて名詞を結合する際に除去する．除去すべき接頭辞および接尾辞の一覧を**表 2** に示す．

以上より，本稿では基本システムに用いた形態素解析によるキーワード抽出「Conventional 手法」，連続する名詞をすべて結合した名詞をキーワードとする「Conjoined 手法」，結合名詞に加え単語 bi-gram から接頭辞/接尾辞の除去を行った語もキーワードとする「Bi-Word 手法」の比較評価を行う．なお，単語 N-gram を用いる手法において N は 2 に限定されるものではない．しかし，いくつかの技術文書に対して予備的に $N = 3$ および $N = 4$ について評価したところ，N を大きくするに従って AP 値が低下した．この傾向は，他の技術文書に対しても同様であると予想できるし，N を 3 より大きくすると，結合名詞の種類が増大し計算時間・メモリ消費も多くなるため，$N = 2$ のみを検討の対象としている．

4.2 概念ベースを用いたクエリ拡張

クエリ拡張とは，検索漏れの低減を目的として，文書検索を行う際に課題文書から抽出されたキーワードに似た概念のキーワードを補助的に加える手法である．この検索漏れが起きる原因として，難波らは学術論文と特許文書において使用される表現方法のゆらぎを指摘している [12]．

そこで，クエリを拡張することにより分類性能の改善を図る．拡張には，対象とする文書セットにおける語の意味的な類似性をより良く反映すると考えられる「概念ベース」を用いることとした．概念ベースとは，文書集合内に出現する個々の語を，文書における他の語との共起頻度に基づいてベクトルとして表現したものである [13]．類似した意味の語は，それが共起する他の語も類似しているという仮説に基づき，共起頻度を統計的に処理して語のベクトルを求めているため，意味が類似した語は互いに近いベクトルを持つことになる．

まず，特許文書集合およびトレーニング用論文集合を用

いて，それぞれの概念ベースを構築する．構築した概念ベースを用いてキーワードどうしの概念ベクトルの距離を計算し，検索キーワードと近い概念のキーワードをクエリに追加する．しかし，単純にこの手法を適用しただけでは性能はほとんど向上しないことが，Shimano らによって報告されているし [14]，筆者らの追実験による評価結果も同様であった．その原因としては，概念ベースの構築に用いた特許文書および論文が広い分野にまたがっていたため，概念ベースの類似判別が対象分野の特徴をうまく反映していないことが考えられる．クエリを拡張する際は，課題である学術論文の持つトピックを推定し，そのトピックに合致した文書集合からクエリに追加するキーワードを選択することが望ましい．

IPC コードは 5 段階の階層構造を有しており，SubClass レベルでは，基本システムでも比較的高い精度が得られている．そこで，基本システムの手法により SubClass レベルで課題の学術論文のトピックを推定し，そのトピックに合致する概念ベースをクエリ拡張に使用することで，分類性能の改善を図る．このために，SubClass ごとに概念ベースを構築し，SubClass レベルで推定した学術論文の IPC コードに基づいて概念ベースを切り替え，クエリの拡張を行う．

具体的な概念ベースの構築およびクエリの拡張の手順は以下のとおりである．まず，特許文書集合に対し，IPC コードの SubClass を基準としてそれぞれに属する文書，およびそれに出現するキーワードの統計情報を集計する．次に，それぞれの IPC コードに属する特許文書集合に出現するキーワードに対し，CF-IDF を用いて重み付けを行う．ここで CF-IDF（Class Frequency – Inverted Document Frequency）とは，それぞれの IPC サブクラスに属する文書集合を統合し，1 つの大きな文書と見なした場合の TF-IDF 値に相当する．すなわち，あるサブクラスにのみ多く出現する特徴的な単語が，そのサブクラスにおいて大きな重みを持つことになる．各文書集合に対し CF-IDF の大きなキーワードを選出し，それらの共起頻度を計算し概念ベースを構築する．

次に，この概念ベースを用いてクエリを拡張する．概念ベースは IPC コードの SubClass ごとに作成されている．クエリ拡張に使用する概念ベースを決定するため，基本システムの手法により SubClass レベルの IPC コードを複数個推定する．これらの IPC SubClass に対応する概念ベースを用いて，もとのクエリのキーワードと類似度の高いキーワードを求めクエリに加える．ただし，クエリに拡張されるキーワードは，その数が多すぎても分類性能を落とす要因になりかねない．そこで，拡張するキーワードを「IPC 偏差値」によりランク付けする．この IPC 偏差値とは SubClass レベルにおけるキーワードの出現の偏りを表現しており，キーワードが多くの IPC SubClass に平均的に出現すればその値は低く，少ない IPC SubClass で局所的に出現すればその値が高くなる．あるキーワード t の IPC 偏差値 $TS(t)$ は以下の式で定義される．

$$TS(t) = 50.0 + \frac{10.0\left(\max(F_{c,t}) - \overline{F_{c,t}}\right)}{\sigma_{(F_{c,t})}} \tag{5}$$

ここで，$F_{c,t}$ は，IPC SubClass c においてキーワード t が出現する特許文書数であり，$\max(F_{c,t})$ は，すべての IPC SubClass の中でキーワード t の出現頻度が最も大きい場合の $F_{c,t}$ の値である．また，$\overline{F_{c,t}}$ は，すべての IPC SubClass にわたる $F_{c,t}$ の平均値であり，$\sigma_{(F_{c,t})}$ は標準偏差である．

概念ベースを用いてクエリに追加される候補となったキーワードは，上記の IPC 偏差値によりランク付けされ，上位から，あらかじめ設定された数だけが実際のクエリ拡張キーワードとして追加される．

この手法では，パラメータとして課題文書に対する IPC コードの候補数（N_{ipc}），拡張キーワードに対する類似度下限（Th_s），拡張キーワード数の元のキーワード数に対する割合（R_{exp}）の 3 つを有する．

4.3 構造を考慮した RSV の算出

特許文書は書誌的事項，発明の概要，特許請求の範囲，発明の詳細な説明，図面の簡単な説明と 5 つの構造要素から構成されており，さらにそれぞれ異なった目的と粒度で内容が記述されている（**表 3**）．

これらの要素は分類するうえで重要度が異なると考えられることから，各要素へ適切な重みを付加することにより分類性能の改善が期待でき，これまでにも，構造を利用す

表 3 特許文書を構成する要素
Table 3 Materials of patent document structure.

発明の名称	発明を最も端的に表す．
発明の概要	従来の技術や技術的な課題，その解決手段が端的に書かれている．
特許請求の範囲	発明が権利を請求する範囲の主張であり，発明の所属する分野や周辺の分野の技術範囲を明確に定めるための技術用語が頻出する．
発明の詳細な説明	当業者が該発明を実施できるように詳細な説明を行う必要があるため，事細かな説明がなされる．そのため最も文章量が多く，内容も広範囲にわたる．
図面の簡単な説明	主に図面中で番号付けされた要素の技術用語が羅列される．この要素の記述は必須ではない．

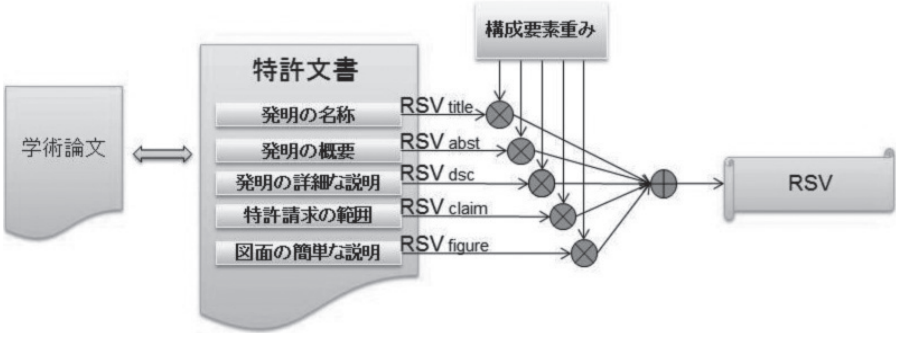

図 2 特許文書の構造を応用した RSV の算出
Fig. 2 Improved RSV method using patent document's structure.

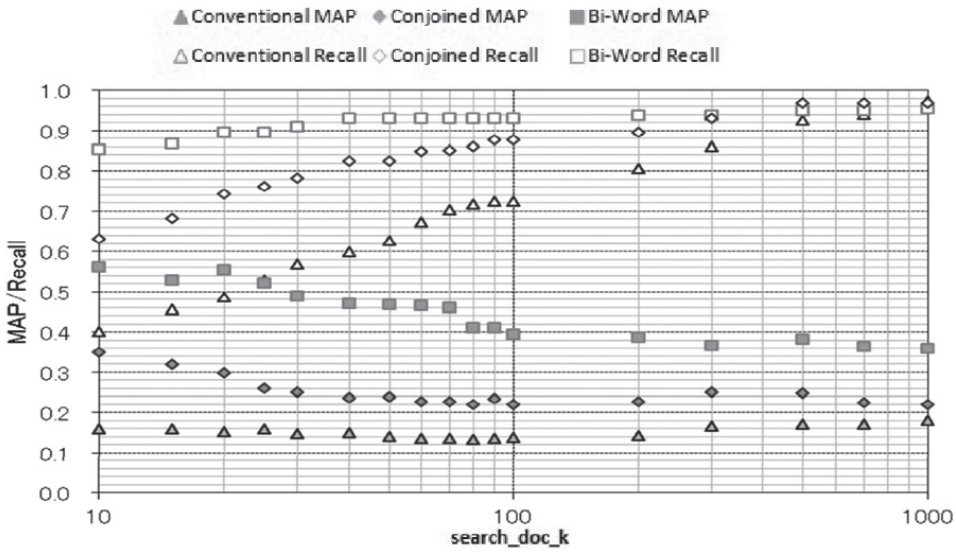

図 3 各キーワード抽出手法による性能
Fig. 3 System performance by each three keywords extraction methods.

る様々な手法が提案されている [5], [15], [16], [17].

本稿では，特許文書を構成する 5 つの要素を個別の文書と見立ててそれぞれ RSV を算出し，それらに重みをつけて線形結合することを考える．ここで，RSV の計算はベースラインシステムと同様に BM25 に基づく．この手法の概要を**図 2** に示す．なお，図中の RSV[title,abst,dsc,claim,figure] は各構造要素に対して算出された RSV を示している．構成要素重みは実験評価に基づいて最適化する．

5. 実験的評価とそれに基づくパラメータの最適化

前章で述べた各要素の高度化に対し，分類性能を実験的に評価し，その性能向上効果を検証した．また，各高度化手法はそれぞれ多くのパラメータを持つ．そこでそれぞれの要素のパラメータを網羅的に変化させて評価することにより，パラメータの最適値を探った．パラメータを実験的に最適化した場合，その最適値の普遍性についても評価が必要である．そのため，NTCIR データセットに含まれる課題 200 件を 66 件と 134 件とに分け，66 件を利用してパラメータの最適値を求め，そのパラメータの基に 134 件の課題を用いて最終的な性能を評価した．

5.1 キーワード抽出

本節では，形態素をキーワードとする Conventional 手法，結合名詞を抽出する Conjoined 手法，さらに単語 bi-gram をキーワードとする Bi-Word 手法ついて評価した結果について述べる．

本システムは，課題文書に対する類似特許文書を検索し，K-NN 法によって IPC を決定しているため，検索件数 k が分類性能に影響するパラメータとなる．そこで，k を 10 から 1,000 まで変化させ，各キーワード抽出手法の MAP 値および再現率を求めた．

その結果を**図 3** に示す．なお，本タスクにおいては，再現率も重要な指標である．課題文書には平均 2.4 個の正解の IPC コードがあるが，これに対し少なくとも 2 個の正解を含む出力が得られるべきと考え，再現率が 0.85 以上となる領域を比較の対象とした．

Conventional および Conjoined 手法は文書収集数 k がそれぞれ 300 件および 70 件以下の場合十分な再現率を得られなかったのに対し，Bi-Word 手法は文書検索件数 k が 10 件

表 4 各キーワード抽出手法の最適値
Table 4 Optimal performance of each keyword extraction method.

	MAP	Recall	k
Conventional	0.178	0.972	1,000
Conjoined	0.249	0.967	500
Bi-Word	0.553	0.894	20

表 5 クエリ拡張の最適化および評価
Table 5 Optimized query expansion and evaluation.

	Th_s	N_{ipc}	R_{exp}	MAP
クエリ拡張なし		—		0.553
クエリ拡張あり	0.95	5	20%	0.560

であっても再現率で 0.851 を得ている．また，Conventional 手法および Conjoined 手法と比較して，Bi-Word 手法がより高い MAP 値を得ていることが分かる．

各キーワード抽出手法で最も MAP 値が高い点の評価値・パラメータを**表 4** に示す．

このように，Bi-Word 手法が最も良好な分類性能を達成しているため，以降の各要素の評価および最適化においてはキーワード抽出に Bi-Word 手法を用いることとし，文書検索件数 k は 20 とする．また，以降の高度化手法では再現率が低下することは考えにくいため，評価指標として MAP 値のみを用いることにする．

5.2 分野を考慮した概念ベースによるクエリ拡張

前章で述べたとおり，この拡張手法は，パラメータとして拡張のために推定する IPC SubClass の個数（N_{ipc}），そして拡張キーワードの類似度下限（Th_s），拡張キーワード数の元のキーワード数に対する割合（R_{exp}），の 3 つを有する．そこで，各パラメータを N_{ipc} を 1〜5，Th_s を 0.5〜1.0，R_{exp} を 0〜50%の範囲で変化させ評価した．最も MAP 値が高かったときの値を**表 5** に示す．

MAP 値が 0.7 パーセントポイント向上しているが，これは小さな値であり，結果としてクエリ拡張では性能の向上はあまり得られないといえる．

先行研究では個々の形態素をキーワードとして抽出しているが，MAP 値にして 0.4060 から 0.4402 と 3.42 パーセントポイントの向上が見られ，クエリ拡張により一定の効果が得られている [5]．本キーワード拡張手法においても，キーワード抽出に Conventional 手法を用いた場合を別途評価したところ，MAP 値が 0.179 から 0.195 へと，精度の絶対値は低いものの，比較的大きな性能向上が得られた．すなわち，個々の単語をキーワードとして抽出している場合には，クエリ拡張は性能を向上させるが，Bi-Word 手法に対しては，あまり効果がないという結果になった．

表 6 構造要素の単独で使用した場合の評価結果
Table 6 Evaluation results for each structure element.

material		MAP
全構造要素		0.553
発明の名称	(title)	0.532
発明の概要	(abst)	0.500
特許請求の範囲	(claim)	0.470
発明の詳細な説明	(desc)	0.538
図面の簡単な説明	(fig)	0.423

表 7 構造要素の重みを網羅的に変化させた場合の評価
Table 7 Evaluation with exhaustive parameter setting of structure weights.

	weight			MAP
	title	abst	desc	
全構成要素		—		0.553
構造要素別重み付け	1	3	2	0.582

5.3 構造を考慮した RSV の算出

前述のとおり，特許文書は 5 つの要素で構成されており，それらが分類性能に与える影響はそれぞれ異なると考えられることから，構造要素ごとに最適な重みをつけることで分類性能の改善が期待できる．

まず個々の構造要素の分類性能への影響を調べるために，各要素を単独で使用した場合の分類性能を評価した．その結果を**表 6** に示す．表中の最上段は全構造要素で 1 つの文書を構成するとした場合の性能（ベースライン）である．

表 6 から分かるように，個々の要素を単独で用いた場合には，どの要素を用いてもベースラインよりも性能が低くなっている．しかしながら，性能を低下させる幅は要素によって異なり，“特許請求の範囲” および “図面の簡単な説明” は特に大きく性能を下げている．これら要素は，類似特許文書検索において適切な文書が検索されるのを妨げる要因となっていると考えられる．そこで，これらの構成要素を取り除き，残った構成要素の重みを最適化して，さらなる分類性能の改善を図る．すなわち，“発明の名称”，“発明の概要” および “発明の詳細な説明” を用い，これら 3 つの構成要素に対し，重みを 1 から 5 まで網羅的に変化させて評価して，重みの最適な組合せを探った．これによって得られた，最も分類性能が高かったときの MAP 値および構造要素の重みを**表 7** に示す．

表 7 より分かるように，特許文書を構成要素に分解してそれぞれの RSV に重み付けすることで，全要素を 1 つの文書として扱う場合よりも高い MAP 値が得られた．また，その際の最適な重みは，(“発明の名称” : “発明の概要” : “発明の詳細な説明”) = (1 : 3 : 2) であった．

5.4 全課題データを用いた総合評価

前節までは，NTCIR から提供された 200 課題のうち 66

表 8 各要素の最適化に基づく性能
Table 8 System performance with preliminary evaluation and all task evaluation.

高度化要素	MAP	
	66 課題（closed test）	134 課題（open test）
基本システム	0.178	0.199
キーワード抽出（Bi-Word）	0.553	0.481
クエリ拡張	0.556	0.481
構造要素重み付け	0.582	0.494

課題を利用してパラメータの最適化を図った．この最適値を用い，残りの 134 課題に対して評価した（オープンテストに相当）．その結果を**表 8** に示す．

まずキーワード抽出手法として Bi-Word 手法をとることにより性能が劇的に向上した．クエリ拡張では，分類性能にほとんど改善が見られなかった．特にオープンテストでは，MAP 値の向上がまったく得られなかった．最後に，構造要素によって RSV 算出の重み付けによっても性能向上が得られた．

以上の結果より，キーワードの抽出手法に関しては，接頭辞・接尾辞が除去された単語 bi-gram を利用することで，分類性能が大きく向上することが明らかとなった．さらに特許文書の構造を RSV に応用することで，さらなる分類性能の向上が見られ，そのとき，最も重要な構造要素は“発明の概要” であることが分かった，

パラメータ最適化に用いた 66 課題に対する MAP 値が 0.582 であるのに対し，オープンテストの 134 課題に対する MAP 値は 0.494 とかなり低い値であった．基本システムでは，それぞれ 0.178，0.199 と，むしろ 134 課題に対する精度が高いことから，本稿で求めた最適値は，必ずしも普遍的な最適値であるとはいえないが，かなり良いパラメータであるといえる．

本稿の手法で得られた MAP 値は，NTCIR において過去に得られた最高値 0.4539 [3] よりも高くなっている．本稿で提案した手法のうち特にキーワード抽出を Bi-Word 方式とすることが精度向上に大きく貢献している．4.1 節でも述べたとおり，一般の文書に対してはこのようなキーワード抽出は性能を悪化させる場合も多く，向上する場合でもそれほど劇的ではない．特許文書の記述形態が一般文書とかなり異なっていることが，本手法を有効としている要因と推察される．

6. まとめ

本稿では NTCIR ワークショップにおける学術論文分類タスクに則り，類似文書検索部にいくつかの要素に改善を施して分類性能の向上を図った．改善手法の分類性能への影響を実験的に明らかにするとともに，それら改善手法が持つパラメータの最適値を求めた．

改善手法としては，(1) 特許文書に多用される長い熟語に着目し，接頭辞・接尾辞を除去し結合名詞とその単語 bi-gram を用いるキーワード抽出法，(2) 検索漏れを低減させるための，分野別に構築した概念ベース利用したクエリの拡張，(3) 特許文書を構成要素に分割し重み付けした RSV 計算を検討した．

個々の改善策について，パラメータを変化させて分類性能の特性を評価し，最適なパラメータを求めた．キーワード抽出手法としては，単語 bi-gram を用いる手法（Bi-Word 手法）が最適であり，類似特許文書検索において上位 20 件を分類に用いるだけで十分に良い精度が得られることが分かった．クエリ拡張は，単純に概念ベースを利用する手法とは異なる分類別概念ベース手法を考案したが，この手法によっても，これまでと同様にほとんど性能向上が得られなかった．構造要素を考慮した RSV 計算では，“発明の名称”，“発明の概要”，“発明の詳細な説明” の RSV を 1 : 3 : 2 の重みで線形結合した場合に最も良い分類性能が得られた．最終的に，再現率を低下させることなく，MAP 値を 0.178 から 0.494 に向上させることができた．

謝辞 本稿ではテストコレクションに NTCIR-8 のデータを利用させていただいた．作成者諸氏に感謝する．さらに，形態素解析処理に形態素解析システム MeCab を利用させていただいた．開発者の京都大学情報学研究科―日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクト諸氏に感謝する．

参考文献

[1] 特許庁：技術分野別特許マップについて，特許庁（オンライン），入手先 ⟨http://www.jpo.go.jp/shiryou/s_sonota/tokumap.htm⟩（参照 2012-03-06）．
[2] Nanba, H., Iwayama, M., Fujii, A. and Hashimoto, T.: Overview of the Patent Mining Task at the NTCIR-7 Workshop, *Proc. NTCIR-7 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, *Proc. NTCIR-8 Workshop Meeting*, pp.293–302 (2010).
[4] Cover, T.M. and Hart, P.E.: Nearest Neighbor Pattern Classification, *IEEE Trans. Information Theory*, Vol.13, No.1, pp.21–27 (1967).
[5] Mase, H. and Iwayama, M.: NTCIR-7 Patent Mining Experiments at Hitachi, *Proc. NTCIR-7 Workshop Meeting*, pp.365–368 (2008).
[6] Mase, H. and Iwayama, M.: NTCIR-8 Research Paper Classification Experiments at Hitachi, *Proc. NTCIR-8 Workshop Meeting*, pp.345–347 (2010).

[7] Gang, J., Qi, K., Jian, Z., Xiaolin, W., Cong, H., Hai, Z. and Bao-Liang, L.: Multiple Strategies for NTCIR-8 Patent Mining at BCMI, *Proc. NTCIR-8 Workshop Meeting*, pp.303–308 (2010).
[8] Robertson, S.E., Walker, S., Jones, S. and Hancock-Beülieu, M.M.: Okapi at TREC-3, *Proc. 3rd Text REtrieval Conference* (*TREC-3*), pp.109–126 (1995).
[9] Lewis, D.D.: Feature Selection and Feature Extraction for Text Categorization, *Proc. Workshop on Speech and Natural Language*, pp.212–217 (1992).
[10] Schütze, H., Hull, D.A. and Pedersen, J.O.: A Comparison of Classifiers and Document Representations for the Routing Problem, *Proc. 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.229–237 (1995).
[11] Tan, C.-M., Wang, Y.-F. and Lee, C.-D.: The Use of Bigrams to Enhance Text Categorization, *Information Processing and Management*, Vol.38, No.4, pp.529–546 (2002).
[12] 難波英嗣，竹澤寿幸：2 種類の翻訳システムを用いた学術論文の特許分類体系への自動分類，情報処理学会論文誌データベース，Vol.2, No.3, pp.76–86 (2009).
[13] Schütze, H. and Pedersen, J.O.: Information Retrieval Based on Word Senses, *Proc. 4th Annual Symposium on Document Analysis and Information Retrieval* (1995).
[14] Shimano, T. and Yukawa, T.: An Automated Research Paper Classification Method for the IPC system with the Concept Base, *Proc. NTCIR-7 Workshop Meeting*, pp.379–384 (2008).
[15] Kim, J.-H., Huang, J.-X., Jung, H.-Y. and Choi, K.-S.: Patent Document Retrieval and Classification at KAIST, *Proc. NTCIR-5 Workshop Meeting* (2005).
[16] Fujino, A. and Isozaki, H.: Multi-label Patent Classification at NTT Communication Science Laboratories, *Proc. NTCIR-6 Workshop Meeting*, pp.381–384 (2007).
[17] Cao, G., Nie, J.-Y. and Shi, L.: NTCIR-7 Patent Mining Experiments at RALI, *Proc. NTCIR-7 Workshop Meeting*, pp.347–350 (2008).

鈴木 克典

2010 年長岡技術科学大学工学部電気電子情報工学課程卒業．2012 年同大学大学院工学研究科電気電子情報工学専攻修了．2012 年 4 月より NEC ソフト株式会社．

湯川 高志 （正会員）

1987 年長岡技術科学大学大学院工学研究科電気電子システム工学専攻修了．同年日本電信電話株式会社入社．2002 年長岡技術科学大学助教授，2012 年より同大学教授．テキストに基づく知識処理システム，特許情報処理システム，知的学習支援システム等の研究に従事．博士（情報学）．電気電子情報通信学会，人工知能学会，日本教育工学会，日本セキュリティマネジメント学会，IEEE 各会員．