特許テーマ分類方式の提案とその評価実験

間瀬久雄[†] 辻洋[†]
絹川博之[†] 石原正博[††]

特許庁における特許審査期間の短縮および審査業務の効率向上を実現するには，出願特許をいかに迅速かつ適切に分類するかが課題の1つとなる．計算機による分類の自動化を実現するためには分類知識が不可欠であるが，特許のように分野が広範囲でカテゴリ体系が大規模である場合，人手で分類知識を構築・保守するには莫大なコストを要する．本論文では大量の分類済み特許データと，各カテゴリの適用範囲を文章で規定した分類マニュアルデータからそれぞれ計算機で抽出したキーワードを統合し，各カテゴリを特徴付けるキーワード集合からなる分類知識を自動生成する方式を提案する．本方式におけるキーワード抽出は，特許明細書の文書構成の特徴を活かすべく，特定フィールドのみに着目した抽出方式および特許文章の構文的特徴を踏まえた重み配分方式を特徴とする．本方式に基づき最大約31万件の特許公報データから分類知識を自動生成し，新規特許公報データを実存する38の上位カテゴリ，2,815の詳細カテゴリに自動分類する評価実験を行い，それぞれ最高96.0%，82.8%の正解率を得た（カテゴリを3種類ずつ付与した場合）．また分類知識保守の観点から，分類知識作成に必要十分な教師データ量を検証する実験を行った結果，約1,000件/カテゴリが必要なことを確認した．実験結果より，分類作業支援システムまたは自動振り分けシステムとして実用化できる見通しを得た．

Automatic Patents Categorization and Its Evaluation

Hisao Mase,[†] Hiroshi Tsuji,[†] Hiroshi Kinukawa[†]
and Masahiro Ishihara[††]

This paper presents keywords-based patents categorization and discusses its simulation study. Patent categorization by computers is helpful for rapid assignment of applied patent documents to appropriate examiners. Then, the classification knowledge is essential for the computers to automatically assign categories to the query documents. We propose a classification knowledge generation method, which extracts keywords that characterize the particular category from a lot of patent documents and a classification guide texts defining the scope of each category. We also propose keyword extraction and ranking method based on the structure of patent documents and the syntactics of the sentences. We did experimental simulation using maximum 310,000 training patent documents. The maximum classification accuracy was 96.0% (38 categories) and 82.8% (2,815 subcategories) when three categories are assigned to each document. We also evaluated how much training data is necessary from the viewpoint of classification knowledge maintenance. The results show that approximately 1,000 patent documents per each subcategory was necessary to classify most correctly and effectively. The results of this simulation strongly encouraged us to develop a patent classification system to support the category assignment work.

1. はじめに

計算機の普及とネットワーク基盤の整備により，各種の情報が電子化された形で流通するようになった．特許出願についても平成2年12月より電子出願が施行されており，年間約37万件もの特許が出願されている．

特許庁では，発明分野・内容に応じて出願特許をカテゴリに分類している．カテゴリは，公知例検索や出願動向分析に不可欠な属性である．しかし，出願件数が膨大であるのに加え，分類が専門家による手作業で行われているため，毎年多大な作業時間と人件費を費やしている．特許審査期間の一層の短縮化が叫ばれている現在，計算機による特許自動分類へのニーズは高まる一方である．

さて，計算機による分類の自動化には分類知識が不

† 株式会社日立製作所システム開発研究所
Systems Development Laboratory, Hitachi Ltd.
†† 財団法人工業所有権協力センター
Industrial Property Cooperation Center

可欠である．しかし，特許の適用分野は広範囲に渡っており，すべての分野を網羅した分類知識を構築，保守するためには莫大なコストがかかることが予想される．

上記課題を解決すべく著者らは，特許庁の外郭団体である（財）工業所有権協力センターからの委託研究として特許自動分類の研究を行っており，本論文では分類知識を自動作成する方式について提案する[1),2)]．すなわち，大量の分類済み特許データと，各カテゴリの適用範囲を文章で規定した分類マニュアルデータからそれぞれ自動抽出したキーワードを統合し，各カテゴリを特徴付けるキーワード集合からなる分類知識を全自動作成する方式を提案する．またキーワード抽出では，特許文書構成の特徴を活かすべく特定フィールドのみに着目した抽出方式，および特許文章の構文的特徴を踏まえた重み付け方式を提案する．

文書分類に関しては，多くの研究成果が報告されている[3),4)]．Hayesら[3)]は，ルールベースで新聞記事を自動分類するシステムを実用化している．しかし，分類知識作成には専門家の介在が必要であり，分野が広範囲である特許分類に適用するには多大なコストが必要となる．

分類知識の自動生成に関しては，文章の統計情報に基づく手法が多く提案されている．湯浅ら[4)]は名詞の出現頻度および共起関係から分類知識を自動生成している．共起情報はキーワードの多義解消に役立つが，処理に用いるキーワード数に制約があるため，専門用語が頻出する特許の分類に適用するのは精度面・性能面で難しい．

余田ら[5)]は，単語の出現頻度に加えて特許文書の構造的・構文的特徴を加味したキーワード抽出を行い，類似特許を出力するシステムを試作した．テキスト間のキーワード照合により類似した特許を認定するが，一テキストから抽出できる情報量には限りがある．したがって本論文で述べるカテゴリ分類に限っていえば，複数テキストから抽出したカテゴリ別の統計データを用いてテキスト対カテゴリのキーワード照合を行う方が，高精度かつ柔軟な分類ができると考えられる．また，同一内容のものが少ない（発明の新規性が問われる）という特許の性質も，テキスト間のキーワード照合の精度を下げる一因になる．さらに，テキスト間のキーワード照合による類似度計算は処理時間がかかるため，過去の大量の特許から類似特許を特定するのは性能面で問題がある．

藤井ら[6)]は，同一段落内の単語の出現状況に応じて同一キーワードを別キーワードとして扱うことにより多義解消する方式を提案している．我々の方式では，多くのカテゴリにまたがって出現するキーワードは多義語（あるいはキーワードとなりえない不要語）と見なす．すなわち，そのキーワードだけでは分類されるべきカテゴリを特定できない単語であると判断し，重要度を比較的低く抑えることにより，結果的に多義語による分類精度の低下を吸収している．

以下，2章では特許明細書の構造的特徴について述べる．3章ではこの特徴を踏まえたキーワード抽出方式，分類知識生成方式，および分類方式について述べる．4章では，本方式の評価実験結果について報告する．

2. 特許明細書の構造的特徴

本章では，分類対象となる特許明細書の構成および特許自動分類の観点から見た特徴について整理する．

平成9年6月現在，特許庁に電子出願される明細書は以下の構成要素（以下，フィールドと呼ぶ）からなる．

【発明の名称】【請求項】【発明の属する技術分野】【従来の技術】【発明が解決しようとする課題】【課題を解決するための手段】【発明の実施の形態】【発明の効果】【図面の簡単な説明】【符号の説明】【図面】【要約】

我々は，新聞記事が「出来事」を記述した文章であるのに対し，特許明細書は「もの」または「もの」に対する「処理」を記述した文章であると考えている．したがって，発明対象となる「もの」あるいは「処理」を端的に表す単語（名詞，動詞）が存在すると考えている．これら重要なキーワードを含む可能性が高いフィールドとして，発明内容を端的に記述した【発明の名称】（タイトル）および発明として権利化したい内容を記述した【請求項】（クレーム）があげられる．

また，【発明の属する技術分野】には，その発明がどの分野で適用されるかが具体的に記述されるため，カテゴリ特定に有効なキーワードを含んでいることが多い．

さらに，上記フィールドのうち，【請求項】および【発明の属する技術分野】に記述される文は特定の構文を有することが多い（3章参照）．これらの構文情報の利用により，重要なキーワードを容易に取得できる．

一方，【発明の実施の形態】（実施例）では，発明を実現する具体的方法が詳細に記述されている．この中には多くの有効なキーワードが含まれている反面，ノイズとなるキーワードも多く含まれている．このフィールドは文章量が多いうえに，有用な構文的特徴が見ら

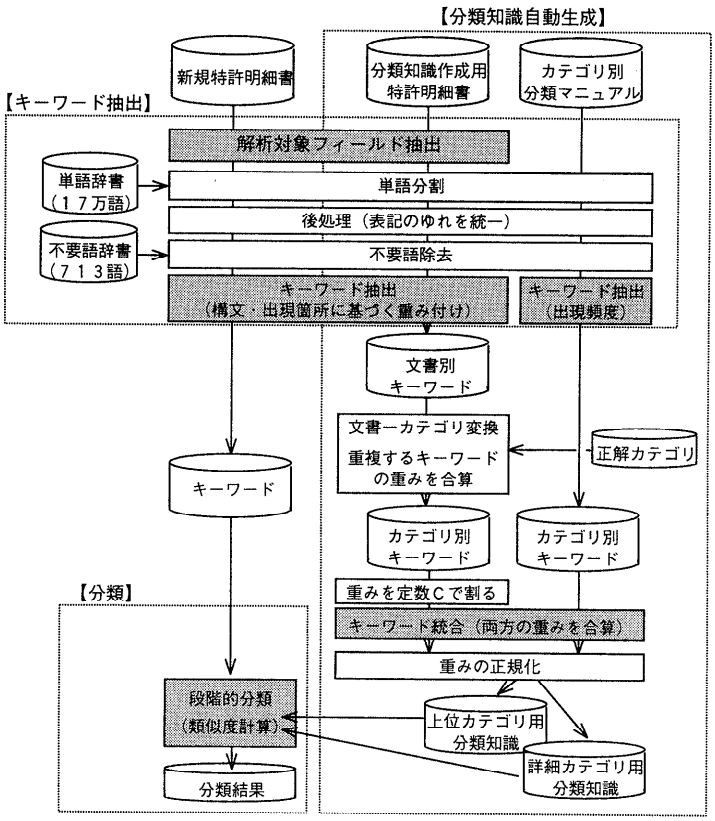


図１ 特許自動分類の処理の流れ
Fig. 1 Process flow of automatic patent categorization.

れないので，キーワード抽出は難しい．

【要約】は発明内容を数百文字でまとめたものであり，有効なキーワードも含まれているが，有用な構文的特徴が見られないこと，キーワードの多くが【発明の名称】【請求項】【発明の属する技術分野】に重複して出現していることから，有効なキーワードを抽出しづらい（500件の特許明細書による予備調査では，要約中のキーワードの約69.0%が上記３フィールドに出現した）．

3. 特許分類アルゴリズム

本章では，前章で述べた特許明細書の特徴を踏まえた分類アルゴリズムについて詳述する．図１に特許分類の処理の流れを示す．特許分類は，特許明細書からのキーワード抽出，分類知識生成，分類の３フェーズからなる．

3.1 特許明細書からのキーワード抽出

次の３つの特徴を有する抽出方式を提案する．

（1）解析対象フィールドの限定

前章で述べた特許明細書の各フィールドの特徴を踏まえ，【発明の名称】【請求項】【発明の属する技術分野】をキーワード抽出の解析対象フィールドとする．ただし【発明の属する技術分野】については，ノイズキーワード除去の観点から，「本発明」「本願」「この発明」「発明は」の４表現のいずれかで始まる最初の一文のみを解析対象とする（500件の特許明細書による予備調査では，99.4%が上記を満たす文を含んでいた）．このように解析箇所を限定することにより，分類処理時間を高速化できる．

（2）構文情報に基づくキーワードの重み付け

上記３フィールドのうち，【請求項】および【発明の属する技術分野】については，その記述方法が比較的定型である．【請求項】では，次の定型文で記述される場合がほとんどである（500件の特許明細書による予備調査では，71.4%について下記構文を満たしていた）．

『[Aにおいて,」Bを備えたことを特徴とするX』

また，【発明の属する技術分野】では，次の構文およびこれに類する構文で記述されることがほとんどである（500件の特許明細書による予備調査では，97.6%につ

表1 キーワード重み付け
Table 1 Weigt assignment for keywords.

項番	重み配分条件	重みの値
1	【請求項】に記述された文の末尾が「AのB」なる名詞句である場合の単語A	5
2	【請求項】に記述された文の末尾の名詞（句）で，項番1以外の単語	3
3	【請求項】に記述された文において，構文「Aを特徴とするX」のAに相当する名詞句を構成する単語	1
4	【請求項】に記述された文に出現する単語	2
5	【発明の名称】に出現する単語	3
6	【発明の属する技術分野】における「本発明」「本願」「この発明」「発明は」の4表現のどれかで始まる文において，「Aに関する」「Aに関した」「Aに関し」「Aに関して」「Aについて」「Aのための」「Aのために」「Aにおける」のどれかの構文を満たす名詞句Aを構成する単語	3
7	【発明の属する技術分野】において「本発明」「本願」「この発明」「発明は」の4表現で始まる文に出現する単語	2
8	項番1または2，項番5，項番6の3条件をすべて満たす単語	10
9	項番1または2，項番5，項番6の3条件のうち，2条件のみを満たす単語	7
10	項番4，項番7を満たす単語	4
注：上記複数の項番を満たす場合，各重みの値の合計値がそのキーワードの重みとなる．第4章の評価実験で用いた分類知識生成の際には，重みが2以下のキーワードを除外した．		

いて，下記構文を満たしていた）．

『本（この）発明は，（～に関し，）～Yに関する』

上記定型文のA，B，X，Yに入る単語（句）は，発明の内容・分野を特定あるいは限定する重要なキーワードと見なすことができる．そこで，これらの位置に出現するキーワードの重みを他の位置に出現するキーワードの重みよりも大きくする．

（3）出現箇所に基づくキーワードの重み付け

また本方式では，「重要キーワードは上記3フィールドの複数箇所に現れる」と仮定している．そこで複数のフィールドに出現するキーワードの重みをそうでないキーワードよりも大きくする．**表1**に，構文および出現箇所に基づくキーワードの重み付け設定方法を示す（4章の評価実験ではこの値を採用した）．これらの重みの値は試行錯誤的に設定する必要がある．

3.2 分類知識生成

本論文における分類知識は，各カテゴリを特徴付けるキーワードとして，どんなキーワードがどのくらいの重要度（重み）で含まれているかを記述しており，各レコードは次のデータ構造を持っている．

『カテゴリコード，キーワード文字列，重み』

（例） 通信，伝送，20
通信，ネットワーク，50
計算機応用，ネットワーク，20
……

上の例で，「伝送」はカテゴリ「通信」を特徴付けるキーワードでその重要度（重み）が20であることを示し，「ネットワーク」はカテゴリ「通信」「計算機応用」の両方を特徴付けるキーワードで，その重要度（重み）はそれぞれ50，20であることを示している．

一般に，教師データからの知識獲得においては，教師データの選定方法を考慮する必要がある．我々は，特許分類知識を生成する際に使用する教師データとして，分類済みの特許明細書のほか，各カテゴリの適用範囲を文章で規定した分類マニュアルが有効であると考えた．

特許のように分野が広範囲でカテゴリが大規模の場合，カテゴリの適用範囲をすべて網羅するのに十分な特許明細書データを収集・選別することは困難である．分類マニュアルは，特許明細書データで網羅できなかった範囲に関するキーワードを分類知識に補足追加する役目を持ち，分類精度を向上できる（4章参照）．

ここで問題となるのは，これら異なる種類の教師データから抽出されたキーワードをどのように統合すべきかである．特許明細書から抽出されたキーワードには構造的特徴を踏まえた重み付けがなされる．一方，分類マニュアルは構造化されていないので，そこから抽出されたキーワードの重み付けは出現頻度に基づいてなされる．したがって，このように異なる観点から配分された重みを統合して1つの分類知識を生成する必要がある．

図2に2種類の教師データから表1の重み付け方法によって抽出されたカテゴリ別のキーワードの重みの分布を示す．図2の横軸の数値は，各々の特許明細書（マニュアル）から抽出されたキーワードに付与された重みを，カテゴリ別かつキーワード別に合計した値である．特許明細書から抽出したキーワードの重みの分布と，分類マニュアルから抽出したキーワードの重みは，重みの大きさは違うものの分布傾向は類似していることが分かる．そこで本方式では，特許明細書から抽出されたキーワードの重みを単純にある定数Cで割ることによって補正した後に重みを合算すること

で統合後のキーワードの重みを配分することにした．ただし定数Cの値は経験的に設定する必要がある（4章参照）．

統合後のキーワードの重みには大小のばらつきが出る．このばらつきはキーワードの重要度の違いによる以外にカテゴリ別の教師データ量の偏りによるところが大きい．教師データから抽出されたキーワードをカテゴリ別にまとめるときにその重みを合算するので（図1参照），教師データ量が多いカテゴリのキーワードの重みは必要以上に大きくなってしまう．

そこで**図3**に示すように，各カテゴリごとに重みの分布幅を一定にし（正規化），重みの価値がカテゴリによらず同一になるように重みをマッピングすることにより，上記問題を解決する．正規化には偏差値の概念を導入する．すなわち各カテゴリについて，ある偏差値のしきい値に対応する重みの値を求め，その値が上限となるように重みを修正した後，各キーワードの重みを1から100の間にマッピングする．このとき，最大値に対する相対比は保存される．この結果，すべてのカテゴリのすべてのキーワードの重みは1から100の間に再配分される（最小の重みを持っていたキーワードは重み1に，上限値以上の重みを持ったキーワードは重み100に正規化される）．

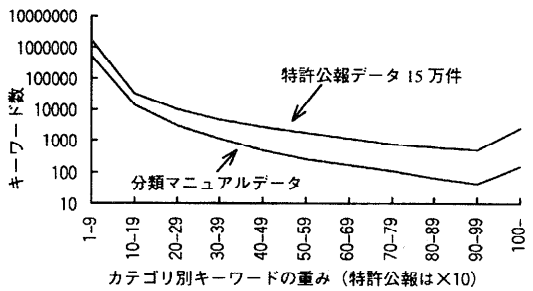


図2 2種類の教師データから抽出されたキーワードの重みの分布
Fig. 2 Distribution of keyword weights from training tests.

3.3 分　　類

3.3.1 カテゴリ別の類似度計算

新規特許文書を分類する場合，分類知識作成時と同一の方法でキーワード抽出および重み付けを行う．そして抽出されたキーワードと分類知識中のキーワードとの照合をカテゴリ単位で行い，各カテゴリに対する類似度を計算し，類似度の高いカテゴリを出力する．

本方式では，次式によりカテゴリの類似度を計算する．

$$S(i) = \Sigma s(i,j) \qquad (j = 1,2,\ldots,n)$$

$$s(i,j) = W(j) \times \sqrt{(w(i,j)/\Sigma w(k,j))} \qquad (k = 1,2,\ldots,m)$$

ここで，$S(i)$はカテゴリiの類似度，$s(i,j)$は新規特許明細書から抽出されたキーワードjに対するカテゴリiの類似度，nはキーワードの種類数，$W(j)$はキーワードjの重み，$w(i,j)$は分類知識中のカテゴリiにおけるキーワードjの重み，$w(k,j)$は分類知識中のカテゴリkにおけるキーワードjの重み，mはカテゴリ数を示している．上式から次の性質を導くことは容易であろう．

(1) 新規特許のキーワードの重みが高いと類似度も高い．

(2) 分類知識のキーワードの重みが高いと類似度も高い．

(3) 多くのカテゴリに現れるキーワードは類似度を低くする．

また上式では，平方根演算子によって類似度を補正している．あるキーワードを含んでいるカテゴリには類似度が加算されるが，この平方根がない場合，類似度は上式の$(w(i,j)/\Sigma w(k,j))$の値に比例する．しかし，直感的にはキーワードを含んでいるカテゴリが重要であるのだから，含んでいないカテゴリとの間の類似度の得点差は大きくすべきである．逆に，大変重要なキーワードであり，上式の$(w(i,j)/\Sigma w(k,j))$の値が1に近い場合，どれも重要とされるのであるから，

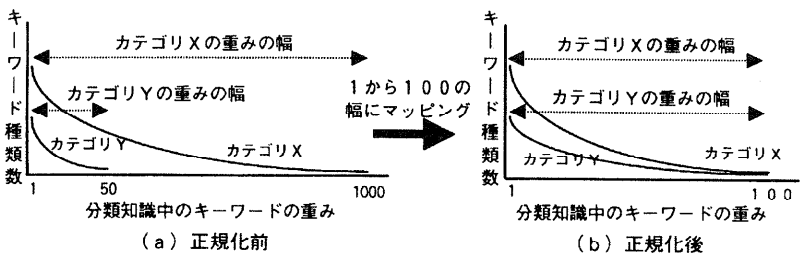


図3 キーワードの重みの正規化の概要
Fig. 3 Overview of keyword weight mapping.

得点差は小さくしてもよいであろう．そこで我々は，類似度計算において上記の性質を近似的に備えている平方根演算子を施して補正している．

3.3.2 段階的分類

特許のように広範囲でカテゴリ数が多い場合，カテゴリをいきなり特定するよりも，まずカテゴリを少数に絞り込んだ後に詳細分類することにより，結果にノイズが含まれるのをある程度防ぐことができると考えられる．そこで，カテゴリ体系が階層的であることを前提として，上位カテゴリの分類結果を詳細カテゴリに反映させる段階的分類方式を以下で提案する．

まず新規特許明細書を上述した方法により上位カテゴリに分類する．次に詳細カテゴリに分類する際に，上位カテゴリで算出された類似度の相対値を詳細カテゴリの類似度計算結果に乗ずることにより，上位カテゴリ分類で上位にランクされたカテゴリに属する詳細カテゴリが分類結果の上位に出力されやすくする．すなわち，次の式により詳細カテゴリの類似度を算出する．

$$S(I) = SU(i) \times \Sigma s(I,j) \qquad (j = 1,2,\ldots,n)$$
$$SU(i) = su(i)/\Sigma su(k) \qquad (k = 1,2,\ldots,m)$$

ここで，$S(I)$ は詳細カテゴリ I の類似度，$SU(i)$ は詳細カテゴリ I の類似度に乗ぜられる係数，$s(I,j)$ は新規特許明細書から抽出されたキーワード j に対する詳細カテゴリ I の類似度（前述の計算式に等しい），n はキーワードの種類数，$su(i)$ は上位カテゴリ i の類似度，$su(k)$ は上位カテゴリ k の類似度，k は上位カテゴリ数を示している．

別の方式として，上位カテゴリ分類結果から上位カテゴリを絞り込み，他を足切りするという方式も考えられる．上位カテゴリの分類精度が100%に近いほどこの方式は有効であるが，そうでない場合，上位カテゴリで誤分類してしまうと詳細カテゴリでは必ず誤分類してしまう．したがって，上位カテゴリの精度に応じてどちらかが良いかを選択するのが最善であろう．

4. 分類精度評価実験

本章では，3章までで述べた方式に基づいて行った，分類精度評価実験方法および結果について報告する．

4.1 実験で使用したデータ

分類体系として，実存する上位カテゴリ（38種類）およびその詳細カテゴリ（2,815種類）を用いた．**図4**にカテゴリの一部を示す．分類知識作成用教師データおよび評価データとして，特許公報32万件（平成5～8年）を用意した．各公報には専門家により上記詳細カテゴリが平均1.9個（上位カテゴリは1.5個）付与されている．

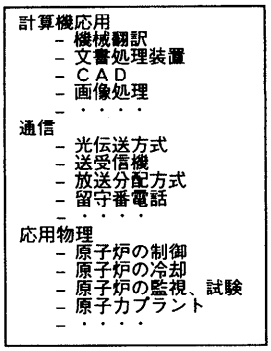


図4 カテゴリの一部
Fig. 4 Sample of categories.

なお，実験ではキーワードとして，名詞，動詞語幹（サ変動詞含む），単位語（「mol」「Hz」など分野固有のものが多い），アルファベット列，カタカナ列，その他の未知語を抽出した．また，キーワードとなり得ない単語（不要語）として，「発明」「特許」など713語を用意した．

4.2 評価尺度について

分類精度を算出する尺度として，再現率と適合率が有名である[7]．再現率が高いほど分類の漏れが少なく，適合率が高いほど分類のノイズが少ない．両者の間には一般にトレードオフの関係がある．

しかし，本実験では次の尺度を正解率として用いる．

$$[\text{上位 } N \text{ 位の正解率（\%）}] = \frac{[\text{上位 } N \text{ 位の中に正解を 1 つ以上含む公報件数}]}{[\text{公報件数の合計}]}$$

すなわち，システムが分類結果として出力した上位 N 個のカテゴリの中に，専門家が分類したカテゴリが「少なくとも1つ以上」含まれている場合，その公報は正解であるとし，正解公報件数の割合を正解率とする．この評価尺度は，人手でカテゴリ分類する作業のうちの一部分でも計算機で補うことができれば，運用形態次第で作業効率が向上するとの認識からきている．

4.3 実験内容および実験結果

本論文では，**表2**に示す4種類の実験について述べる．

[実験1] 教師データの種類の違いによる分類精度検証

特許公報，分類マニュアルから自動生成した分類知識を用いて分類したときの精度を比較検証した．分類マニュアルのみ，公報のみ（1万件から31万件まで増加させる），両方を統合した場合，の3パターンについて分類精度を測定した．また，分類知識作成用の特許公報を増加させたときの分類精度の変化を検証する

表 2 分類評価実験内容一覧
Table 2 Contents list of experimental simulation.

項番	実験内容	教師データ	評価データ	備考
1	・教師データの種類の違いによる分類精度を比較 ・必要十分な教師データ量の把握	・特許公報1-31万件 ・分類マニュアル ・上記2種類を統合	特許公報 3850件	3種類の教師データでの分類精度を比較
2	重み付けに関するパラメータのチューニングによる分類精度向上を検証	特許公報15-31万件と分類マニュアルを統合	特許公報 3850件	3種類のパラメータを最適化
3	大量の評価データによる分類精度評価	特許公報15万件と分類マニュアルを統合	特許公報 約16万件	3種類のパラメータを最適化
4	特許自動振り分けシステムの実現可能性検証	特許公報15万件と分類マニュアルを統合	特許公報 約16万件	3種類のパラメータを最適化

表 3 特許広報および分類知識に関する統計データ
Table 3 Statistical data on patent documents and classification knowledge.

項目	特許公報15万件	分類マニュアル	両方を統合
抽出されたキーワードの種類数	78,444種類	51,036種類	98,749種類
生成された分類知識レコード数	1,706,679	553,030	2,027,323
1カテゴリあたりのキーワード種類数	606種類	196種類	720種類
最も多くのキーワードをもつカテゴリのキーワード種類数	5,479種類	1,968種類	5,789種類
最も少ないキーワードをもつカテゴリのキーワード種類数	0種類 (26カテゴリ)	1種類	3種類
注：詳細カテゴリ（2815種類）に関する統計データ			

ことにより，分類知識作成に必要十分な教師データ量を把握することができる．これは分類知識保守の観点（データの収集作業・解析時間の効率化）から重要なデータとなる．なお，評価データとして新規公報3,850件を用いた．

表 3 にキーワードに関する統計データを示す．また，実験結果を**図 5** に示す．図5に示すように，2種類の教師データを統合した場合が最も正解率が高く，公報のみの場合に比べて2.2%（78.6－76.4）から10.5%（69.9－59.4）向上している．公報から抽出できなかったキーワードを分類マニュアル中のキーワードが補完していると考えられる．一方，分類マニュアルのみの場合の分類精度は他の2つに比べて大きく低下した．この理由として，データ量が少ないこと，記述形式が特許明細書の記述形式と異なること，分類マニュアルではカテゴリの範囲を規定する抽象的/集合的な単語が使われやすい，の3点があげられる．したがって，分類マニュアルはあくまで補完的なデータとして位置づけるべきであろう．

また，公報件数が15万件付近で分類精度がほぼ頭打ちになっている．必要十分な教師データ量は，カテゴリの定義やその数に依存するが，精度のピークが公報件数が15万件のときとした場合，1カテゴリあたり約1,000件（150,000 × 1.9 ÷ 2,815 ＝ 約1,000）の教師データ（公報）があれば必要十分である．

[実験 2] 重みのチューニングによる精度向上検証

重みをいろいろチューニング（最適化）することに

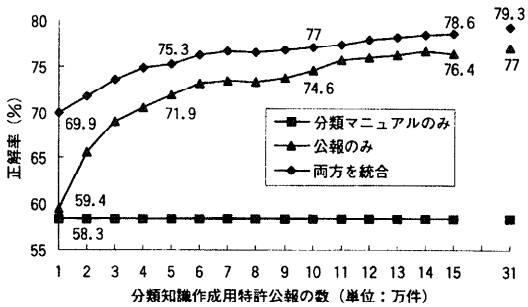


図 5 分類知識作成用教師データの違いによる精度比較 (実験1)
Fig. 5 Comparison of classification accuracy (Exp. 1).

よって分類精度を向上できる．重みのチューニング方法としては，(1) カテゴリ別のキーワードの重みの上限値の設定，(2) 特許公報中のキーワードと分類マニュアル中のキーワードの統合方法の最適化（図1における定数 C の値の設定），(3) 分類知識の重みの正規化の適用（3.2節）の3種類があると考えられる．

本実験では，上記3種類のチューニングパラメータを変化させたときの分類精度を検証することにより，分類知識のチューニングの効果を検証した（本論文では各パラメータの最適値における精度のみについて述べる）．

実験結果を**表 4** に示す．特許公報15万件と分類マニュアルを統合した分類知識を用いた場合で比較すると，上記3種類のチューニングにより分類精度を最大5.3%（60.3－55.0）向上できた（表4の項番5と項番

表 4 3種類の重みチューニングパラメータの最適化による精度向上（実験 2）
Table 4 Classification accuracy based on tuning of three parameters (Exp. 2).

項番	分類知識作成方法		詳細2815カテゴリ分類精度（%）				上位38カテゴリ分類精度（%）		
	教師データ種類	最適化するパラメータ	上位1位	上位3位	上位5位	上位10位	上位1位	上位2位	上位3位
1	分類マニュアルのみ	最適化しない	35.4	57.8	66.7	77.1	73.5	87.4	92.3
2	分類マニュアルのみ	重みの上限を最適化	37.6	58.5	67.3	77.4	73.5	87.4	92.3
3	特許公報15万件のみ	最適化しない	54.6	75.9	82.8	89.3	80.5	91.1	95.0
4	特許公報15万件のみ	重みの上限を最適化	55.8	77.2	83.2	89.6	81.4	91.4	95.3
5	上記2種を統合	最適化しない	55.0	76.8	83.8	90.0	80.7	91.6	95.3
6	上記2種を統合	重みの上限を最適化	57.2	78.6	85.0	90.5	81.5	91.9	95.5
7	上記2種を統合	重みの上限と統合方法を最適化	58.0	79.4	85.6	91.0	82.7	92.7	95.8
8	上記2種を統合	重みの上限と統合方法を最適化し、さらに正規化を施す	60.3	80.9	87.2	92.4	82.4	92.5	95.6
9	特許公報31万件と分類マニュアルを統合	重みの上限と統合方法を最適化し、さらに正規化を施す	61.6	82.8	88.6	93.2	83.3	93.2	96.0

表 5 3種類の重みチューニングパラメータの最適値（表 4 項番 9 の場合）
Table 5 Optimum values of three parameters (#9, Table 4).

チューニングパラメータ	上位38カテゴリへの分類	詳細2815カテゴリへの分類
カテゴリ別キーワードの重みの上限値（分類マニュアル）	1000	30
カテゴリ別キーワードの重みの上限値（31万件特許明細書）	30000	1500
キーワード統合方法（定数Cの値）	20	15
重みの正規化（偏差値のしきい値）	正規化せず	200

8の詳細分類上位1位の正解率を比較した）．特に正規化の効果が大きい（最大 2.3%（60.3 − 58.0）の精度向上）ことを確認した．上記チューニングパラメータの最適値は，教師データの量に依存するので，今回用いた最適値が必ずしもつねに最適であるとは限らないが，最適化により精度向上を図れることが分かった．また，上位 38 カテゴリ分類においては，正規化の効果がないことを確認した．この原因としては，上位 38 カテゴリ別の教師データの量に格差がそれほど見られず（最多と最少の格差が 4.0 倍にとどまった），正規化前後で重みの分布に違いが見られなかったことがあげられる．なお，分類精度が最良であるとき（表 4 の項番 9）の各チューニングパラメータの最適値を**表 5** に示す．

[実験 3] 大量の評価データによる評価実験

上記 2 種類の実験で用いた評価データは 3,850 件であるが，2,815 カテゴリへの分類ということを考慮すると量が少ない．そこで本実験では 16 万件の公報を評価データとすることにより，本方式が精度的に安定しているかを検証する．分類知識は実験 2 で最適化した後のもの（2 種類の教師データを統合したもの）を用いた．

実験結果を**表 6** に示す．分類精度は評価データが

表 6 大量の評価データによる精度評価（実験 3）
Table 6 Simulation using 166,323 evaluation data (Exp. 3).

	詳細2815カテゴリへの分類精度（%）				上位38カテゴリへの分類精度（%）		
評価データの量	上位1位	上位3位	上位5位	上位10位	上位1位	上位2位	上位3位
3,850件	60.3	80.9	87.2	92.4	82.4	92.5	95.6
166,323件	60.7	80.4	86.4	92.0	81.3	91.8	95.1

3,850 件のときに比べて最大 0.8%（87.2 − 86.4），上位カテゴリについては最大 1.1%（82.4 − 81.3）の精度低下にとどまり，精度的に安定していることが分かった．

[実験 4] 特許自動振り分けシステムの実現可能性検証

本実験では，特許を全自動で漏れなく上位カテゴリに振り分けられるかという観点から分類精度を検証した．実験 3 の上位カテゴリ分類結果を別の評価尺度，すなわち，上位 3 個のカテゴリの中に，専門家が分類した正解カテゴリがいくつ含まれているかによって本方式を評価した．

実験結果を図 **6** に示す．棒グラフの面積が件数に比例していることに注意されたい．全体の 64.5%の公報がただ 1 つの正解カテゴリを持つが，そのうち 94.2%については，分類結果として出力された上位 3

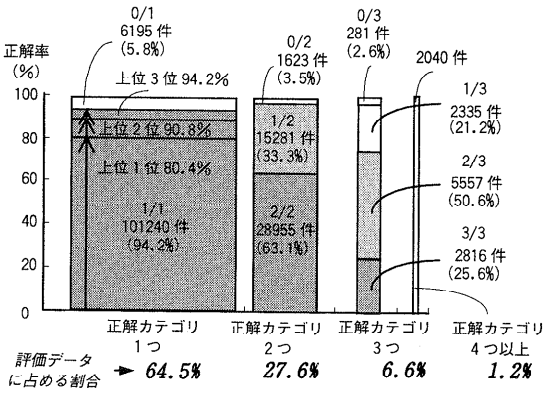


図 6 上位カテゴリの分類精度 (実験 4)
Fig. 6 Classification accuracy of upper category (Exp. 4).

個のカテゴリの中にその正解カテゴリを含んでいた．また，全体の 80.0%の公報（濃い網掛けの部分）については，上位 3 個のカテゴリの中に全正解カテゴリを含んでいた．

複数のカテゴリに分類されるべき特許について，すべてのカテゴリを上位に出力することは，本論文のアプローチでは難しい．なぜなら，各カテゴリを特徴付けるキーワードが対象特許文書から必ずしも十分な数だけ得られるわけではないからである．たとえば，特許公報 A がカテゴリ X，Y の 2 つに分類されるべきだとした場合，X，Y に関連するキーワードがそれぞれ十分な数だけ抽出できれば，X，Y の両方に分類することは本方式でも可能である．しかし，その公報の内容の大部分がカテゴリ X に関するものであり，カテゴリ Y に関する内容は補足的なものである場合，その公報に含まれるキーワードは X に関するものが多くなり，カテゴリ Y に関するキーワードは相対的に少なくなる．その結果，分類結果として Y を上位に出力することは困難となる．したがって，すべてのカテゴリを出力できるようにするためには，少数派のカテゴリ（Y）に対応するキーワードの認定および重み付けについて改良する必要があるだろう．

4.4 誤分類事例の分析

誤って分類された公報 100 件について，その原因を手作業で分析した結果を**表 7** に示す．重要キーワードの照合もれ（分類知識中の特定カテゴリにキーワードが含まれない）と，重み付けの大小による誤分類という 2 種類の原因だけで全体の約 9 割を占めた．キーワード照合もれを改善する方策としては，シソーラスや共起語を用いたキーワード展開により柔軟なキーワード照合を行うことが考えられる．これらのデータは文脈に基づく単語の意味付けが可能である反面，ノイズ単語の増加による精度低下の恐れがあるので，各カテゴリ別のキーワード分布の特徴とシソーラス/共起語の特徴を融合したような方式が必要となろう．また，重み付けの大小による誤分類を改善するためには，キーワード別の重みのチューニングが必要である．運用時における重み付けの学習方式は重要な研究課題となるであろう．

表 7 誤分類結果の分析
Table 7 The causes of the wrong results.

項番	原因	該当件数
1	重要キーワードの照合もれ	48
2	キーワードの重みの大小	42
3	複合語表現による照合もれ	9
4	単語辞書の語彙不足	6
5	不要語除去による照合もれ	4
6	構文・解析箇所による重み付けの失敗	2
7	表記の差異（発明者の記述ミス含む）	2
8	文章が短くキーワードが不足	1
9	その他	1
注：100件対象，複数可		

4.5 処理速度

特許自動分類の処理の流れは図 1 に示すとおりである．特許明細書 1 件にかかる処理時間は約 9 秒である（ワークステーション 3050RX/330T, 105 MIPS, メモリ 192 MB，15 万件の特許明細書を教師データとした場合）．このうち約 7 割が文章の単語分割処理に使われる．一方，分類知識作成に要する時間は，教師データとして使われる特許明細書の件数にほぼ比例しており，約 7 秒/件である．したがって，数十万件規模の特許明細書から知識ベースを作成しようとすると多大な時間がかかるが，新規特許明細書を分類するときにキーワード抽出結果をログとして残しておくことにより，それを分類知識更新時に利用できるので，実運用上は分類知識作成に時間はかからないと考える．

5. おわりに

電子出願特許を対象とした自動分類方式について述べた．分類済み特許データと分類マニュアルデータから分類知識を全自動作成する方式を提案した．また，特許明細書の文書構成の特徴を活かすべく，特定フィールドのみに着目したキーワード抽出方式および特許文章の構文的特徴を踏まえた重み配分方式を提案した．

本方式に基づき最大約 31 万件の特許公報データから分類知識を自動生成し，新規特許公報データを実存する 38 の上位カテゴリ，2,815 の詳細カテゴリに自動分類する評価実験を行い，それぞれ最高 96.0%，82.8%の正解率を得た（カテゴリを 3 種類ずつ付与した場合）．また分類知識保守の観点から，分類知識作

成に必要十分な教師データ量を検証する実験を行った結果，約1,000件/カテゴリが必要なことを確認した．

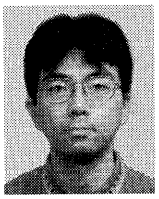
自動分類精度が100%になることは不可能であるので，分類結果の正誤をユーザが簡単に判定せしめるためのユーザインタフェースあるいはデータ提供について検討することが今後の必須課題となろう．

謝辞 本研究の機会を与えていただくとともに貴重なご助言をいただいた（財）工業所有権協力センターの方々に感謝します．また，本研究を推進するにあたり多大なご支援およびご協力をいただいた（株）日立製作所公共情報事業部の細矢良智氏，甲谷和也氏，藤田恵美理氏に感謝します．

参考文献

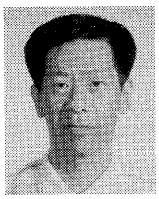
1) Mase, H., Tsuji, H., Kinukawa, H., Hosoya, Y., Koutani, K. and Kiyota, K.: Experimental Simulation for Automatic Patent Categorization, *Proc. Advances in Production Management Systems '96*, pp.377–382 (1996).
2) 間瀬久雄，森本由起子，辻　洋，絹川博之：テキスト分類支援ツールFLUTEの開発（1）—機能と構成，第52回情報処理学会全国大会論文集，3-303 (1996).
3) Hayes, J. and Weinstein, S.P.: CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories, *Proc. 2nd Annual Conference on Innovative Applications of Artificial Intelligence*, pp.1–5 (1990).
4) 湯浅夏樹：大量文書データ中の単語共起を利用した文書分類，情報処理学会論文誌，Vol.36, No.8, pp.1819–1827 (1995).
5) 余田直之，湯村　武，西田行輝：言語情報に基づく検索，Info-Tech94講演論文集，pp.138–146 (1994).
6) 藤井洋一，鈴木克志，今村　誠，高山泰博：共起情報を利用した文書の自動分類，情報処理学会研究報告，NL118-16 (1997).
7) Salton, G.: *Automatic Text Processing: The Trans-formation, Analysis, and Retrieval of Information by Computer*, pp.313–326, Addison-Wesley (1989).

(平成 9 年 9 月 30 日受付)
(平成 10 年 4 月 3 日採録)



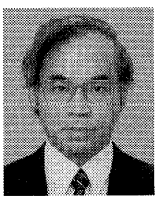
間瀬　久雄（正会員）

1965年生．1988年名古屋大学工学部電気工学科卒業．1990年同大学院工学研究科情報工学専攻修士課程修了．同年（株）日立製作所入社．以来，自然言語処理，文書処理の研究に従事．現在，同社システム開発研究所関西システムラボラトリに所属．人工知能学会会員．



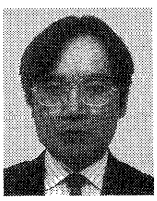
辻　　洋（正会員）

1978年（株）日立製作所入社．システム開発研究所にて，意思決定支援システム，知識ベースシステム，グループウェアシステムの研究・開発に従事．現在，関西システムラボラトリ主任研究員．この間，カーネギメロン大学客員研究員，システム制御情報学会編集委員，電子協専門委員等歴任．情報処理学会編集委員．博士（工学）．技術士（情報処理部門）．



絹川　博之（正会員）

1947年生．1970年東京大学理学部数学科卒業．同年，（株）日立製作所入社．以来，漢字・日本語情報処理システム，仮名漢字変換，自動インデクシング，日本語文書処理，自然語インタフェース，自然言語処理技術の研究開発に従事．現在，同社システム開発研究所研究主幹．理学博士．1996年10月より東京工業大学大学院総合理工学研究科客員教授を兼任．昭和61年度情報処理学会論文賞，平成7年度関東地方発明表彰発明奨励賞，各受賞．電子情報通信学会，言語処理学会，ACL各会員．



石原　正博

1953年生．1976年東京大学工学部船舶工学科卒業．同年特許庁へ入庁．現在，工業所有権協力センター企画部在職．