情 報 学 基 礎 70−2
デジタル・ドキュメント 38−2
（２００３．３．２８）

語とカテゴリの結合の強さを考慮した特許自動分類の検討

米森 力† 原 正巳†

本稿では，語とカテゴリの結合の強さを利用することで，学習文書に付与されているカテゴリのうち重要なカテゴリを決定し，分類する手法を提案する．提案手法では，まず，語とカテゴリとが一対一に対応する辞書を用いて各々の学習文書に付与された複数のカテゴリから学習文書中の重視すべきカテゴリを選別しておく．新たに分類対象文書にカテゴリを付与する際には，重視すべきカテゴリの重要度を高くすることで，より適切な分類の付与を試みた．特許明細書を対象として，k-NN と比較実験した結果、提案手法の有効性を確認した．

Automatic Patent Categorization Using Degree of Linkage Between Phrase and Category

Chikara Yonemori† and Masami Hara†

This paper describes a method of categorizing a text by using an information about a predefined linkage list between phrase and category. In this method, it gives a category describing contents best in each training text in advance. The category can be decided when the text includes an important word in the linkage list. To categorize a new text properly, the category in each training text descibed above is given a higher priority than any others. According to the evaluation results, this method is proved to be over 2 points more effective than k-NN in categorizing texts.

1. はじめに

近年，情報の電子化が進み，データベースには大量のデータが格納されるようになった．この状況は特許についても例外ではなく，毎年 40 万件あまりの特許が出願されている．このような膨大な情報に迅速にアクセスするために，特許庁のホームページ内にある特許電子図書館をはじめ，多くの有料サービスが存在する．

検索における多様なニーズに応え，また文書をその内容に応じて分類するため，特許には特有なカテゴリ体系である国際特許分類（IPC）が付与されている．現在，IPC は専門家が手作業で付与しているが，膨大な数の出願に対して，適切な IPC を付与することは困難であり，特許審査の遅延の一因となっている．

この問題を解決する技術としてテキスト自動分類技術がある．テキスト自動分類の導入によって，大量の文書を短時間で処理できるようになり，作業量の低減や迅速な情報公開にもつながる．

従来のテキスト分類の代表的な手法には，パターン認識の分野で研究が進められてきた SVM(support vector machine)[2)14)] を用いる手法や，ベイズ学習を使った手法[10)] や決定リスト[13)] などが挙げられる．これらの手法は，あらかじめカテゴリが付与された学習文書からカテゴリとの特徴を学習した後，分類対象文書との類似性が高いカテゴリを分類対象文書に付与することで分類を実現している．

従来手法のうち現実的な時間で学習や分類が可能であり，直感的に理解しやすいテキスト分類手法の一つとして k-Nearest Neighbor(k-NN) がある[15)]．k-NN は，あらかじめ分類が付与された学習文書から，分類対象文書に類似している k 件を抽出し，k 件に付与されたカテゴリから分類対象文書のカテゴリを決定する手法である．一般的に，1 件の学習文書に複数のカテゴリが付与されていることも少なくないが，k-NN では，各カテゴリには分類対象文書に対して同等の類似性を持つものとして扱われる．しかし，本来文書に付与されている複数のカテゴリが等しく内容を反映しているとは考えにくく，重視すべきカテゴリとそれ以外のカテゴリが存在すると考えられる．したがって，これらを区別して扱うことができれば，より適切なカテゴリを付与できる可能性がある．本検討では，同一

† 株式会社 NTT データ 技術開発本部
Research and Development Headquarters, NTT DATA Corporation.

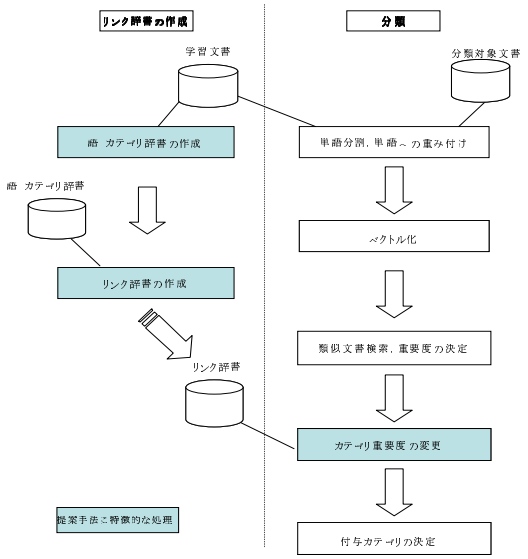

図 1 処理概要
Fig. 1 Overview the procedure.

学習文書に付与されている複数のカテゴリのうち，重視すべきカテゴリと副次的なカテゴリを選別し，重要度に応じた値をカテゴリに与えて分類する方法を提案する．

提案手法の処理概要を図 1 に示す．次節以降では，まず，提案手法の特徴と処理について詳細に説明し，分類実験の結果及び考察を述べる．

2. 提 案 手 法

重要なカテゴリをその他のカテゴリと区別するには，重要カテゴリを特定する必要がある．そこで本検討では学習文書中の語に着目した．学習文書中でそれらの語が特定できれば，該当文書の重視すべきカテゴリが明らかとなる．

本節では，重要カテゴリの発見方法を述べた後，分類手順を述べる．

2.1 リンク辞書の作成

辞書作成は，まず，語-カテゴリ辞書を作成し，次に語-カテゴリ辞書を用いて重視すべきカテゴリを決定するという 2 段階の手順をもって行われる．語-カテゴリ辞書の作成では，文書構造化技術[5] を用いて，学習文書から重要語を抽出する．次に，その語と学習文書に付与されているカテゴリとの組に対して，結合の度合いを母比率検定で確認した後，得られた対応表が語-カテゴリ辞書である．

続いて，文書 ID に対する重視すべきカテゴリを得る方法について図 2 を基に説明する．まず，学習文書集合から取り出した 1 つの学習文書から重要箇所を抽出し，これに対して語-カテゴリ辞書に登録されている語が含まれるかどうかを調べる．語が含まれていた場合には，文書 ID と当該語とを組にして，重複なく記憶しておく．以上の処理を学習文書全てに対して行った後，得られた文書 ID と語の組および前述の語-カテゴリ辞書から文書 ID と重視すべきカテゴリの組を得る．文書 ID とカテゴリの組の集合を以降では，リンク辞書と呼ぶ．

2.2 分 類 手 順

分類対象文書を分類するには，まず，学習文書に対して形態素解析をしたのち，得られた単語を素性とした特徴ベクトルを作る．ある文書 d の素性 t の値 w_t を次式で決定する．

$$w_t = \frac{tfidf(t,d)}{\sqrt{\sum_L tfidf^2(t,d)}} \tag{1}$$

ただし，L は学習文書から抽出した単語の異なり語数である．ここで，$tfidf$[8] は次式で表される．

$$tfidf(t,d) = tf(t,d) \times idf(t) \tag{2}$$

$$idf(t) = \log\left(\frac{N}{df(t)}\right) + 1 \tag{3}$$

ここで $tf(t,d)$ は，素性 t が該当文書に現れる頻度であり，$df(t)$ は素性 t の全学習文書に対する出現文書数である．学習文書と同様の方法で，分類対象文書の特徴ベクトルを作る．

次に，分類対象文書に類似した学習文書 k 件を，類似度の高い順に取り出す．(4) 式に類似度の計算方法を示す．

$$sim(A,B) = \vec{a} \cdot \vec{b} \tag{4}$$

ここで，$sim(A,B)$ は分類対象文書 A と学習文書 B の類似度，$\vec{a}, \vec{b}$ はそれぞれ分類対象文書 A，学習文書 B の特徴ベクトルである．

分類対象文書のカテゴリを決定するにはまず，上記の k 件からなる類似文書に対し，類似度を各々の類似文書に付与されたカテゴリの重要度とする．次に，類似文書の文書 ID とリンク辞書を照合し，ID と対応するカテゴリがあれば，そのカテゴリが他のカテゴリよりも高い重要度を持つように r 倍することで重要度の調整を行う．

ここで r は，重視すべきカテゴリの重要度を表すパラメータである．以上の処理を k 件の類似文書全てに適用した後，カテゴリ毎に重要度を合算し，重要度が高いカテゴリを分類対象文書に付与する．

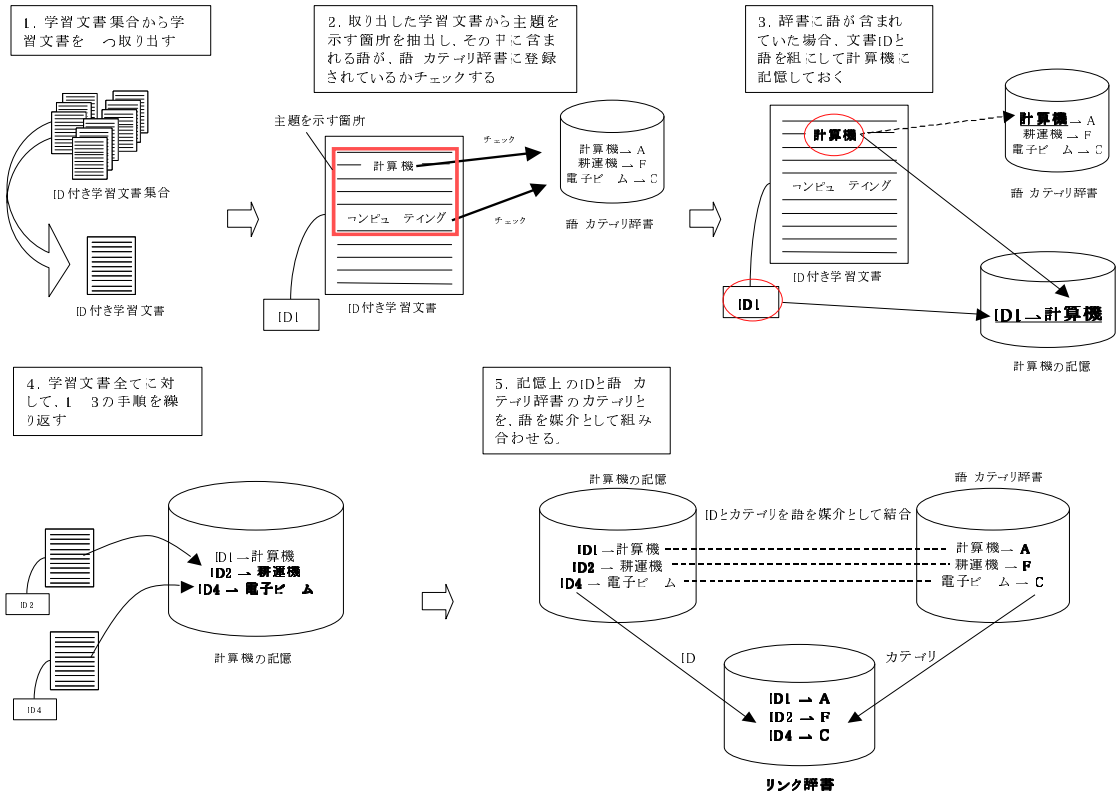

図 2　リンク辞書の作成手順
Fig. 2　Procedure making link dictionary

3. 評価実験

3.1　実験データ

特許明細書は【特許請求の範囲】,【実施例】などの“【”，“】”で囲まれた項目で内容毎に分かれている．

特徴ベクトルの作成には，重要な語の抽出箇所を，間瀬らの論文[1]に基づき,【発明の名称】と【請求項】とし，そこに含まれる文を ChaSen[9] で形態素解析し，名詞，サ変動詞，未知語，アルファベット列，カタカナ列の単語を使用した．

実験用の学習文書には 1995 年の公開公報 95600 件を，分類対象となる評価文書には 1997 年の公開公報から 10000 件をランダムに抽出して使用した．

付与するカテゴリは，5 階層ある IPC の分類体系 (図 3) のうち 2 階層目のクラスと 5 階層目のサブグループとした．クラスはカテゴリを 118 個持ち，サブグループはカテゴリを 57320 個持つ．今回の実験データにおいて，学習文書に付与されているカテゴリが，評価文書に付与されているカテゴリをどの程度網羅できているかを確認したところ，クラスで 91%，サブグループで 62%をカバーしていることがわかった．また，学習文書に付与されているカテゴリの数は，クラスでは平均 1.4 個であり，サブグループでは平均 2.4 個であった．

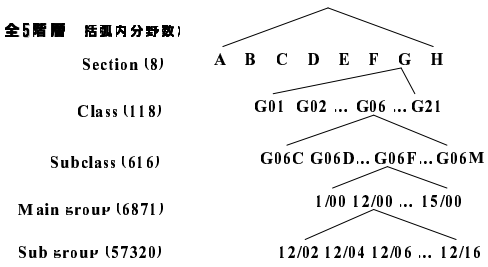

図 3　IPC の分類体系
Fig. 3　Category tree of IPC

3.2　評価指標

分類性能の評価には F 値のマイクロ平均[15] を用いる．F 値は以下の式で定義される．

$$\text{F 値} = \frac{2PR}{P+R} \tag{5}$$

P と R はそれぞれ適合率と再現率 (図 4) であり，以下の式で表される．

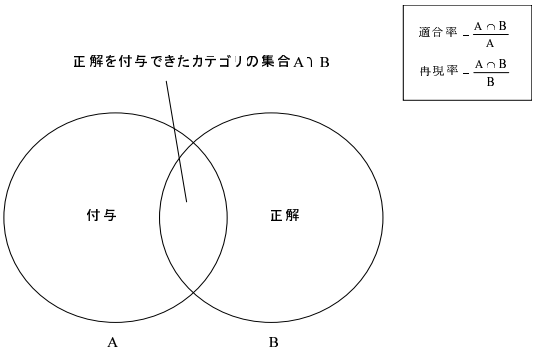

図 4 適合率と再現率
Fig. 4 Precision and recall

$$\text{適合率 } P = \frac{\text{正解数}}{\text{付与カテゴリ数}} \tag{6}$$

$$\text{再現率 } R = \frac{\text{正解数}}{\text{正解カテゴリ数}} \tag{7}$$

適合率は付与したカテゴリが，どの程度正解に合致したかを表す指標であり，再現率は正解をどの程度漏れなく付与できたかを表す指標である．

3.3 実　　験

重要度の調整による正解カテゴリの順位変動を確認するため，重要度調整用のパラメータ r に 2.0 と 4.0 を設定し，付与カテゴリ数を 1 個から 10 個へと段階的に増やし，提案手法と k-NN の精度を比較した．

なお，付与できるカテゴリが規定の個数に満たない場合，可能な数だけカテゴリを付与することとした．k は予備実験で k-NN が高い精度を示した 40 とした．

3.4 実 験 結 果

重要度調整による正解カテゴリの順位変動を，図 5 と図 6 及び表 1 と表 2 に示す．図 5，6 は重要度の高い順にカテゴリを付与したときの F 値の変化を現している．

今回の実験結果では，クラスでは一部で，サブグループでは全体的に，提案手法の F 値が k-NN の F 値よりも高い精度を示しており，学習文書の特定のカテゴリについて重要度を高くする手法が精度向上に有効であることがわかる．

図 5 と表 1 では，付与カテゴリ数が 2 のとき，提案手法の F 値が k-NN の F 値よりも 1 ポイント高い値を示すが，付与カテゴリ数を増やすにつれて，提案手法の F 値は k-NN の F 値とほぼ同一の値となることがわかる．また r を変化させることによる F 値の変化は，ほとんど見られない．

一方，図 6 と表 2 からは，付与カテゴリ数によらず，提案手法の F 値は k-NN よりも高い値を示すことがわかる．特に付与カテゴリ数が 3 のとき，提案手法の F 値は 0.15 を超え，k-NN の F 値よりも 2 ポイント高い．また，調整用のパラメータ r の値を増すにつれて，より高い F 値が得られている．

また，付与するカテゴリ数を増やしていくと，F 値

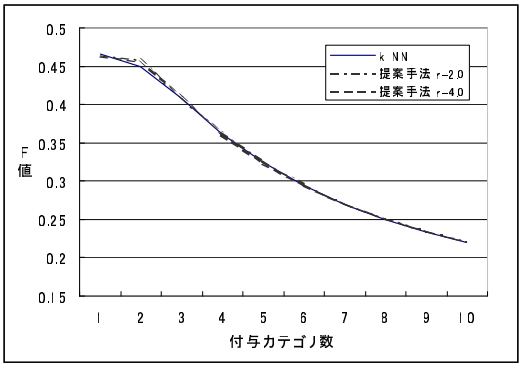

図 5 付与カテゴリ数に対する F 値の変化（クラス）
Fig. 5 F value by category number (class).

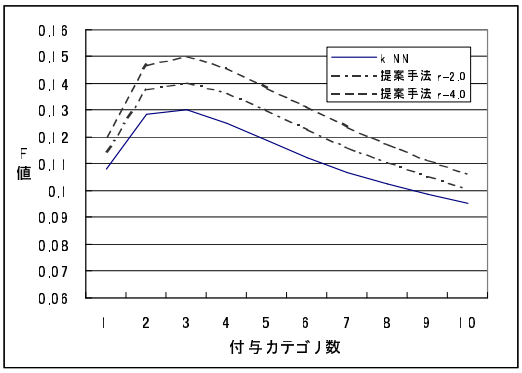

図 6 付与カテゴリ数に対する F 値の変化（サブグループ）
Fig. 6 F value by category number (subgroup).

表 1 実験結果（クラス）
Table 1 Experimental result. (class)

付与カテゴリ数	評価指標	k-NN	r = 2.0	r = 4.0
1	適合率	0.559	0.557	0.554
	再現率	0.398	0.397	0.397
	F 値	0.465	0.464	0.462
2	適合率	0.381	0.394	0.390
	再現率	0.545	0.557	0.552
	F 値	0.449	0.459	0.457
3	適合率	0.300	0.300	0.302
	再現率	0.636	0.637	0.635
	F 値	0.407	0.408	0.410
4	適合率	0.246	0.246	0.244
	再現率	0.690	0.692	0.686
	F 値	0.362	0.363	0.360
5	適合率	0.209	0.210	0.208
	再現率	0.729	0.732	0.724
	F 値	0.325	0.326	0.323

表 2　実験結果 （サブグループ）
Table 2　Experimental result. (subgroup)

付与カテゴリ数	評価指標	k-NN	r = 2.0	r = 4.0
1	適合率	0.184	0.195	0.204
	再現率	0.076	0.081	0.085
	F 値	0.108	0.115	0.120
2	適合率	0.141	0.152	0.162
	再現率	0.118	0.126	0.135
	F 値	0.128	0.138	0.147
3	適合率	0.117	0.126	0.135
	再現率	0.146	0.158	0.169
	F 値	0.130	0.140	0.150
4	適合率	0.100	0.109	0.117
	再現率	0.167	0.182	0.195
	F 値	0.125	0.136	0.146
5	適合率	0.088	0.097	0.102
	再現率	0.183	0.201	0.213
	F 値	0.119	0.130	0.138

は評価文書のカテゴリ数の平均値付近 (クラス: 1.4 個, サブグループ 2.4 個) で最大の値を示した後，減少していくことがわかる．

4. 考　　察

実験でクラスとサブグループに精度の向上度合いに違いが生じた理由について考察する．本検討では，複数のカテゴリが付与されている文書を重要度調整の対象としている．そこで，1 件あたりの付与カテゴリ数が 2 個以上の学習文書の割合を確認したところ，クラスでは 48%であり，サブグループでは 87%であった．重要度調整の対象となる学習文書数がクラスとサブグループとで異なっていたことが，精度向上の度合いに違いが生じた理由と考えられる．

クラスとサブグループとで F 値の最大が得られる付与カテゴリ数が異なるのは，評価文書に付与されている正解カテゴリ数の平均の違いによるものと考えられる．例えば，付与カテゴリが 3 個のとき，サブグループの F 値が最大になるのは，評価文書に付与されているカテゴリ数の平均が 2.4 個であるためと考えられる．仮に評価文書に付与するカテゴリ数を 1 とし，その全てが正解に合致したとしても，再現率の上限が 40%と低くなる．このことが，正解カテゴリ数の平均より少ないカテゴリ数を付与した場合の，精度の低下につながったと考えられる．また正解カテゴリ数よりも多くのカテゴリを付与した場合は適合率が低下するため，F 値も低下することになる．

5. おわりに

従来の k-NN では学習文書に付与されているカテゴリの重要度を考慮せずにカテゴリを決定しているため，重要度が適切に与えられず精度低下の要因となっているとの考えに基づき，学習文書中の重要語と結びつきの強いカテゴリを重要なカテゴリとして，分類対象文書カテゴリの決定に反映させる手法を提案した．また，提案手法が分類精度向上に貢献することを実験によって確認した．特に学習文書 1 件あたりのカテゴリの数が多いほど，提案手法の効果が顕著であることがわかった．

本検討では，語と文書との結合の有無のみに着目して，調整用のパラメータを一意に決定しているが，実際には重要語と文書との結び付き度合は文書毎に異なると考えられるため，文書毎にパラメータの調整が必要と考えられる．パラメータの調整方法として，語-カテゴリ辞書の作成において得られる関連度を反映する方法がある．関連度は語とカテゴリの結合の強さを表していると考えられるため，この関連度を考慮すればより適切なカテゴリを付与できる可能性がある．

また，構文や位置などの表層的な情報を利用して，文書中のキーワードが現れやすい箇所を特定し，文書内容を明確に表す語を抽出することで，文書とカテゴリとの組を増やす方法も考えられる．表層的な情報を利用する方法はテキスト要約[3] やキーワード抽出[4] の分野で研究されており，これらの知見を活かす方向で検討を進めていきたい．

付　　録

A.1　語-カテゴリ辞書の例

重要語	カテゴリ	カテゴリの説明
C 系合金磁石	C22C 38/00	鉄合金, 例. 合金鋼
多層回路基板	H05K 3/46	多重層回路の製造
樹脂成形方法	H01L 21/56	封緘, 例. 被覆

謝辞　本研究を行う機会を与えてくださった浜口技術開発本部長，松本技術開発副本部長，島崎主幹研究員に感謝いたします．また常日頃，有益なコメントを頂くテキストチームの諸氏に感謝いたします．

参　考　文　献

1) 間瀬久雄, 辻 洋, 絹川博之, 石原雅弘: 特許テーマ分類方式の提案とその評価実験, 情報処理, Vol. 39, No. 7, pp. 2207–2216 (1988).
2) 高村大也, 松本祐治: 独立成分分析を用いた文書分類, 自然言語処理, Vol. 143, No. 3, pp. 17–23 (2001).

3) 原 正巳, 木谷 強, 江里口 善生: 特徴的表現を利用した特許抄録作成法の検討, 自然言語処理, Vol. 100, No.14, pp. 105–112 (1994)
4) 原 正巳, 中島浩之, 木谷 強: テキストのフォーマットと単語の範囲内重要度を利用したキーワード抽出, 情報処理, Vol. 38, No.2, pp. 299–309 (1997).
5) 江里口善生, 木谷 強: パターンマッチング手法による名称特定処理の検討, 自然言語処理, Vol. 115, No. 10, pp. 67–73 (1996).
6) 特許庁編: 特許・実用新案 国際特許分類表 〔第6版〕-IPC- (1995).
7) 徳永健伸: 情報検索と言語処理, 東京大学出版 (1999).
8) 長尾 真: 自然言語処理, 岩波書店 (1996).
9) 松本裕治, 北内啓, 山下達雄, 平野善隆, 松田寛, 浅原正幸: 日本語形態素解析システム『茶筌』version 2.0 使用説明書第二版, Information Science Technical Report NAISTIS-TR99012, Nara Institute of Science and Technology (1999).
10) Dumais S. T., Platt J., Heckerman D. and Sahami M.: *Inductive learning algorithms and representations for text categorization*, In Proceedings of ACM-CIKM98, pp. 148–155 (1998).
11) Frakes W. B. and Yates R. Y.: *Information Retrieval* Data Structures & Algorithms, Prentice Hall, New Jersey (1992).
12) Nigam K., McCallum A., Thrun S. and Mitchell T.: *Text Classification from Labeled and Unlabeled Documents using EM*, Machine Learning, 39, pp. 103–134 (2000).
13) Rivest R. L. : *Learning decision lists*, Machine Learning, Vol. 2, pp. 229–246 (1987).
14) Smora A. J., Bartlett P. L., Schölkoph B. and Schuurmans D.: *Advances in Large Margin Classifiers* , MIT Press (2000).
15) Yang, Y.: *An Evaluation of Statistical Approaches to Text Categorization.*, Information Retrieval, Vol. 1, 1-2, pp. 69–90 (2001).