自 然 言 語 処 理 106－3
（1995．3．9）

テキスト分類のためのカテゴリ割り付け戦略

西野文人
nisino@flab.fujitsu.co.jp
富士通研究所 ソフトウェア研究部
〒211 川崎市中原区上小田中1015

テキスト自動分類のための新しいカテゴリ割り付け戦略を提案する．本稿で提案するカテゴリ割り付け戦略は，1つの文書内で各カテゴリの割り付け確率値を比較するのではなく，訓練セットに対して分類処理を実行してみて，各カテゴリの割り付け確率値と実際にそのカテゴリが割り付けられているかどうかの関係（割り付けの信頼度）を調べておき，その信頼度に基づいてカテゴリを割り付けるものである．本カテゴリ割り付け戦略の優位性は特許の自動分類（国際特許分類コード付与）実験を通して確認された．

Category Assignment Strategy for Text Categorization

Fumihito NISHINO

Software Laboratory, Fujitsu Laboratories Ltd.
1015, Kamikodanaka, Nakahara-ku, Kawasaki, Kanagawa, 211 Japan

We propose a new category assignment strategy in automatic text categorization. Traditional systems generally employ category assignment strategies which compare the probability values within a single document. In our approach, we don't use these values directly, rather, we translate the probability values into "believability values" by examining the relations of the probability values and category assignments in the training set. Then, the categories are assigned to the document using the believability values. The superiority of this strategy is confirmed by categorization experiments of patent documents.

1 はじめに

「与えられたテキストがどの分野と最も類似しているかを判定する」技術は，テキスト自動分類として知られており，様々な研究が行なわれている [1, 2, 3, 4, 5, 6, 7, 8, 9]．しかし，その多くはターム抽出や類似度計算の手法についてのものであり，カテゴリ割り付け戦略に目を向けているものは多くない．本稿ではテキスト自動分類のための新しいカテゴリ割り付け戦略について述べる．本稿で提案するカテゴリ割り付け戦略は，1つの文書内で確率値を比較するのではなく，訓練セットに対して分類処理を実行してみて，各カテゴリに対する確率値と実際にそのカテゴリが割り付けられているかどうかの関係（割り付けの信頼度）を調べておき，その信頼度に基づいてカテゴリを割り付けるものである．本カテゴリ割り付け戦略の優位性は特許の自動分類（国際特許分類コード [10, 11] 付与）実験を通して確認された．

2 自動分類の手法

テキスト分類の手法の一つとして，言語情報や知識を利用したアプローチがある．このアプローチでは，潜在的に高精度な分類を行なえる可能性があるが，大規模な知識ベースの構築・維持管理技術はまだ確立されておらず，実用化が困難である．これに対して，手軽に実現可能な手法として，統計的な情報を利用したアプローチがある．

統計手法を利用したテキスト分類の一般的な手順は以下のようになる．

1. 前もって与えられるカテゴリの特徴を内部表現形式（ベクトル，決定木など）に変換する．
2. 分類したい文書の特徴を内部表現形式（ベクトルなど）に変換する．
3. カテゴリの特徴の内部表現と文書の特徴の内部表現との距離を計算する（類似度，確率など）
4. カテゴリを割りつける（分類コードなど）

カテゴリや文書を内部表現にして距離計算を行なうモデルとしては，特徴をベクトルで表現し類似度を計算するベクタモデルと，確率を用いる確率モデルがある．本稿では，このような距離計算を行なったあとで，どのカテゴリを割り付けるかを決定するカテゴリ割り付け戦略に注目することにする．

3 カテゴリの割り付け戦略

3.1 従来の手法

いま文書を3つのカテゴリ (A,B,C) に割り付けることを考えてみる．

	各カテゴリに含まれる確率		
文書番号	カテゴリ A	カテゴリ B	カテゴリ C
100	0.5	0.3	0.2
101	0.6	0.1	0.3
102	0.7	0.1	0.2
103	0.5	0.2	0.3

上の表はそれぞれの文書が与えられた時に，これらの文書がカテゴリ A，カテゴリ B，カテゴリ C に属する確率（以下では確率ということで例を議論するが類似度でも同様である）を示したものである．

カテゴリ割り付け戦略として良く知られているものに K-per-doc および probability threshold という手法がある．

K-per-doc 各文書の確率値の上位 k 個のカテゴリを割り付ける．

例えば $k = 1$ とすると，上の例では文書番号 100, 101, 102, 103 のどの文書に対してもカテゴリ A を割り当てることになる．

この戦略ではすべての文書に対して同じ個数のカテゴリを割り付けることになる．

probability threshold ある閾値以上の確率値をもつカテゴリを割り付ける．

例えば，0.3 以上の確率値をもつカテゴリを割り付けることにすると，文書番号 100 に対してはカテゴリ A と B を，文書番号 101 に対してはカテゴリ A と C を，文書番号 102 に対してはカテゴリ A を，文書番号 103 に対してはカテゴリ A と C を割り付けることになる．

このように，これまでのカテゴリ割り付け戦略は1つの文書内で確率値を比較して割り付けるカテゴリを決定していた．

これに対して，Lewis[6] が，「1つの文書内で確率値を比較するのではなく，そのカテゴリ全体で比較すべきだ」ということを主張し，次のようなカテゴリ割り付け戦略を提案した．

proportional assignment 訓練例におけるカテゴリの確率分布に比例した割り付け（各カテゴリに対して，そのカテゴリでのトップスコアのものから順に，そのカテゴリが出現する比率に応じた数だけ，そのカテゴリを割り付ける）

この戦略によれば，まず各カテゴリがどのくらいの分布になるかを調べておく必要がある．今，例えばカテゴリ A が 25%，カテゴリ B が 25%，カテゴリ C が 50% の割合で付与されているとする．重なり係数を 1.0 とすれば，テストセットは4文書なので，カテゴリ A は $4 \times 1.0 \times 0.25$ から1つの文書にだけ，カテゴリ B も1つ，カテゴリ C は2つの文書に割り付けることになる．それぞれのカテゴリ内での確率値を比較して，カテゴリ A

は一番確率値の高い文書102に，カテゴリBは文書101に割り当てる．カテゴリCは2位までに割り当てるので文書101と文書103に割り当てることになる．

Lewis[6]や岩山[9]の実験では，k-per-doc < Prob. thresholding < Prop. assignmentという結果が得られている．

3.2 信頼度に基づくカテゴリ割り付け戦略

proportional assignment戦略ではテストセットに割り当てられるカテゴリの数の分布が訓練セットと同じであることを想定している．テストセットの文書数が十分多ければ良いが，実際の運用での処理単位では必ずしもテストセットの量が十分あるとは限らないという問題がある．

ここではカテゴリ全体で確率値を比較した時に，カテゴリの出現分布を利用してカテゴリを割り付けるかどうかの閾値を決めるのではなく，以下のような考え方を提案する．

> 確率値（あるいは類似度，距離値など）の値を直接利用するのではなく，その値とカテゴリ割りつけの正解／不正解との関係（信頼度）で判断するべきである．

ここで信頼度を「文書dがカテゴリCに属する確率値$p_c(d)$がx以上の値であった時に実際に文書dがカテゴリCに属する確率$P(d \in C|p_c(d) \geq x)$」と考えることにする．この値は，ベイズの定理を使って以下のように変形することができる．

$$
\begin{aligned}
P(d \in C|p_c(d) \geq x) &= \frac{P(d \in C) \cdot P(p_c(d) \geq x|d \in C)}{P(p_c(d) \geq x)} \\
&= \frac{\frac{F(p_c(d) \geq 0|d \in C)}{F(p_c(d) \geq 0)} \cdot \frac{F(p_c(d) \geq x|d \in C)}{F(p_c(d) \geq 0|d \in C)}}{\frac{F(p_c(d) \geq x)}{F(p_c(d) \geq 0)}} \\
&= \frac{F(p_c(d) \geq x|d \in C)}{F(p_c(d) \geq x)}
\end{aligned}
$$

ここで$F(p_c(d) \geq x)$は標本において文書dがカテゴリCに属する確率値$p_c(d)$がx以上であると判定された文書の数，$F(p_c(d) \geq x|d \in C)$は実際にカテゴリCに属する文書の中でその文書のカテゴリCに属する確率$p_c(d)$がx以上であると判定された文書の数を示す．これらの値は，訓練セットを標本として，訓練セットの各文書dに対して$p_c(d)$を求めて，その値がx以上になる文書の数と，その中で本来カテゴリCが割り当てられるべき文書の数を求めて推定することになる．

信頼度に基づくカテゴリ付与 訓練セットを使って得られたモデルをまず訓練セットに適用して，各文書の各カテゴリに属する確率値を求める．各カテゴリに対して，各文書のそのカテゴリに属する確率と実際にカテゴリを付与すべきかどうかのデータからカテゴリに属する確率値と信頼度との関係を求める．この信頼度がある一定の値を越えるものに対してそのカテゴリを付与する．

実際の実行では，文書dがカテゴリCに属する確率値を信頼度に変換するのではなく，あらかじめ定めた値を越えるような信頼度が得られる確率値の下限を各カテゴリごとに求め，文書dがカテゴリCに属する確率値$p_C(d)$がその値を越えているかどうかでカテゴリ付与の決定を行う．

4 実験

4つのカテゴリ割り付け戦略の優劣を調べるために，特許文書を分類し国際特許分類コード(IPC)を付与する実験を行なった．

4.1 特許分類とは

発明特許の技術内容を記載した公報類は各国のものを総計すると毎年百万件にも達するといわれている．これを活用するために世界共通に利用できる分類として国際特許分類があり，日本では特許・実用新案の公告公報のすべてにこの国際特許分類コードが専門の審査官によって付与されている．

特許の分類の仕方には，技術を応用分野（用途）別に細分化する方法と，技術を機能（固有の性質）別に細分化する方法とがあるが，国際特許分類では機能別分類を基調とし，特殊用途への適用という分野別分類を加味した全体的には両者の折衷的な分類体系になっている．体系の構造は，全技術分野がAからHの8つのセクション（A: 生活必需品，B: 処理操作；運輸，C: 化学；冶金，D: 繊維；紙，E: 固定構造物，F: 機械工学；照明；加熱；武器；爆破，G: 物理学，H: 電気）に分かれており，その中がさらに，クラス(118)，サブクラス(616)，メイングループ(6871)，サブグループ(57320)に分かれるというような木構造になっている．例えば，国際特許分類G06F15/38というのは，セクションが「G 物理学」であり，クラスは「G06 計算；計数」，サブクラスが「G06F 計算の少なくとも一部は電気的に行なわれるデジタル計算機；計算機デジタルデータを取り扱う装置」，メイングループが「G06F15 グループ7/00と少なくとも他の1つのメイングループに含まれる機能の組合せを特徴とするデータ処理装置」，サブグループが「G06F15/38 言語ほん訳のためのもの」を示している．この例でも示したように，各グループははっきりした特徴を持つものだけでなく，グループの中には「他に分類されない…」として示される残余事項分類や「2以上の先行グループからなる…」として示されるような新規な複数の特徴を持つ技術事項や多数の択一的なものからなる技術事項が分類されるグループも存在している．分類コードは1つの特許文書に1つとは限らず，複数の分類コードが付与されているものも多い．

4.2 データ

公開特許実用新案公報(JPO CD-ROM)の1993年の公開番号特開平5-038201から特開平5-084000の特許(CD-ROM 12枚分) 45,749件を利用した(公開番号には欠番がある)．また以下のすべての実験では，特開平5-038201から特開平5-077241までの39,000件を訓練セットとして用い，特開平5-077242から特開平5-084000までの6,749件をテストセットとして用いた．

実験としては，1) 与えられた特許文書を8つのセクションに分類する実験，2) 与えられた特許文書をセクションの下位分類である118のクラスに分類する実験，3)8つの各セクションのそれぞれに対して，そのセクションに属する(IPCの少なくとも1つがそのセクションのコードを持つ)特許文書のみを対象としてセクションの下位分類であるクラスへ分類する実験，4) 出現頻度の高い10のクラスを選び，その各クラスに属する特許文書を下位のサブクラスに分類する実験，5) 出現頻度の高い10のサブクラスを選び，その各クラスに属する特許文書を下位のメイングループに分類する実験，合わせて30の実験を行なった．

各実験に対する下位カテゴリ数(定義上のものではなく訓練セットに実際に現れた数)，実験に用いた訓練文書数，テスト文書数を以下に示す．

	下位カテゴリ数	訓練文書数	テスト文書数
セクションへ	8	39000	6749
クラスへ	118	39000	6749
A	15	3186	210
B	34	9173	1241
C	19	5632	883
D	8	760	128
E	7	1300	333
F	17	3773	645
G	13	13944	2552
H	5	12165	2481
A61	11	987	121
B41	8	1344	265
G01	17	2835	471
G02	3	1603	282
G03	6	2520	459
G06	7	3960	839
G11	2	2218	357
H01	14	5940	1190
H04	10	4052	832
H05	5	1192	230
A61K	12	549	105
B41J	19	923	183
G01N	20	846	150
G02B	14	993	150
G03G	7	1353	229
G06F	9	3674	774
G11B	16	1875	282
H01L	13	3389	600
H04N	8	2277	452
H05K	7	775	139

4.3 モデル

岩山[9]は確率モデルとしてSVMV(Single Random Variable with Multiple Values)モデルを提案し，ベクトルモデルのTF・IDFおよび3つの確率モデルRPI, PRW, CTとの比較実験を行ない，$TF \cdot IDF < RPI < PRW < CT < SVMV$ という結果を示している．今回の実験では，岩山の提案SVMVモデルを採用することにした．このモデルでは，文書 d がカテゴリ C に含まれる確率を以下のようにして求める．

$$
\begin{aligned}
&P(d \in C|D = d) \\
&= \sum_{t_i} P(d \in C|D = d, T = t_i) P(T = t_i|D = d) \\
&= \sum_{t_i} P(d \in C|T = t_i) P(T = t_i|D = d) \\
&= P(d \in C) \sum_{t_i} \frac{P(T = t_i|d \in C) P(T = t_i|D = d)}{P(T = t_i)}
\end{aligned}
$$

ここで，それぞれの項の意味は以下のとおりである．

分類対象テキストの内部表現形式

$P(T = t|D = d)$ 文書 d からタームをランダム抽出した時，それが t である確率．

$$\left(= \frac{\text{その文書におけるターム}t\text{の頻度}}{\text{その文書の総ターム数}} \right)$$

カテゴリの内部表現

$P(T = t|d \in C)$ カテゴリ C の文書からタームをランダムに抽出した時，それが t である確率．

$$\left(= \frac{\text{カテゴリ}C\text{におけるターム}t\text{の頻度}}{\text{カテゴリ}C\text{の総ターム数}} \right)$$

$P(T = t)$ 任意の文書からタームをランダムに抽出した時，それが t である確率．

$$\left(= \frac{\text{全データにおけるターム}t\text{の頻度}}{\text{全データの総ターム数}} \right)$$

$P(d \in C)$ 文書をランダム抽出した時，d のカテゴリが C の確率．

$$\left(= \frac{\text{カテゴリ}C\text{の文書数}}{\text{全文書数}} \right)$$

4.4 ターム抽出

特許文書中の<要約>，<特許請求の範囲>，<発明の詳細な説明>の部分から形態素解析を行なって名詞のみを抽出した．同義語や表記のゆれに関する処理は特に行なわなかった．

タームを抽出した結果は，1文書あたりの異なりターム数としては，30 ～ 1600語，平均は約240語であった． 訓練文書39,000件での異なりターム数は259,314語，全45,749文書では異なりタームは288,546語であった．

4.5 分類システムの評価法

分類の評価はカテゴリ単位の適合率(precision)と再現率(recall)とによって行なった.

再現率 (recall) 付与すべきカテゴリのうちで正しく付与した率

適合率 (precision) 付与したカテゴリのうちで正しく付与した率

5 実験結果

全体を8つのセクションに分類した実験結果(図1)と,118のクラスに分類した実験結果(図2),クラスをサブクラスに分類した実験とサブクラスをメイングループに分類した実験の中から1つずつの結果(図3,図4)を示す.

ほとんどの実験で,岩山の実験と同様に,k-per-docやprobability thresholdよりproportinal assignment手法が良い結果をもたらした.しかし,一部にはproportional assignmentがあまり良い結果を示さないものがあった(例えばクラスA61の下位メイングループへの分類).これは予想どおり,訓練セットにおける下位カテゴリの分布とテストセットにおける下位カテゴリの分布に著しく差があったためである.

今回提案した信頼度に基づく手法は,多くの実験でproportional assignmentより良い結果をもたらした.少なくとも,proportional assignmentを含めた3つの手法より劣ることはなかった.

今回提案した手法では,文書dがカテゴリCに属する確率を信頼度に変換するために,訓練文書のそれぞれに対して各カテゴリに属する確率値を計算することが必要である.これは,訓練文書数×カテゴリ数のオーダーの時間がかかる.しかし,この信頼度を求めるのにそれほど多くの訓練文書数は必要ないであろうし,カテゴリ割り付けの前に1回求めておけば良いだけなので,それほど問題はないであろう.

6 おわりに

本稿では,求まった確率値をそのまま利用するのではなく確率値とカテゴリ割り付けの正解/不正解との関係を求め,それを利用してカテゴリの割り付けを行う手法を提案し,カテゴリの割り付けの精度の点で大きく改善できることを示した.

今回の実験では4万6千件の特許データを利用したが,特許は年に約40万件が公開されており,今回のデータはわずか1か月半分のデータにしかすぎない.特許文書は現在は多くの審査官によってIPCが付与されており,このコード付与作業の支援が望まれている.特許の自動分類という観点からはもっと大量のデータで実験することが必要であろう.

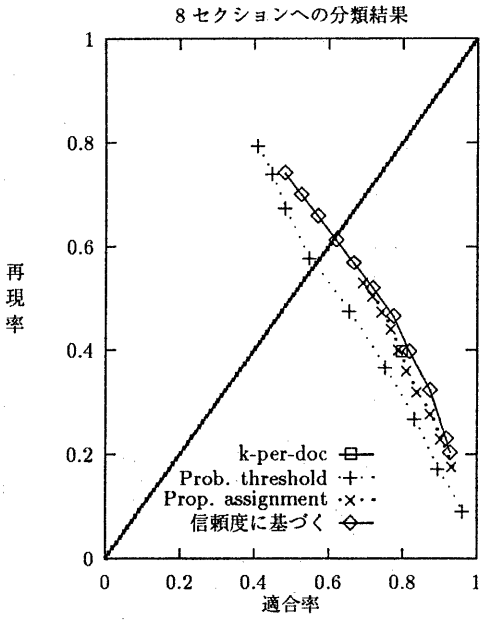

図 1: 適合率－再現率曲線(8セクションへの分類)

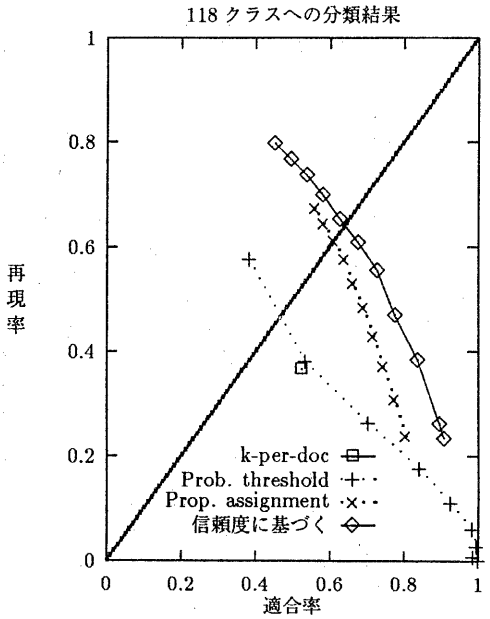

図 2: 適合率－再現率曲線(118クラスへの分類)

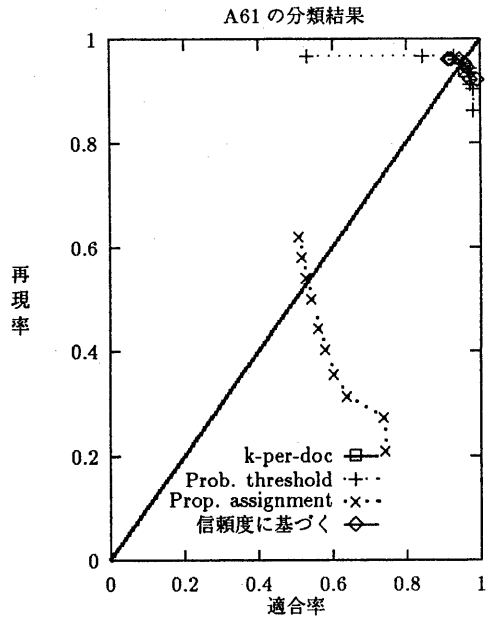

図 3: 適合率－再現率曲線（クラスからサブクラスへの分類）

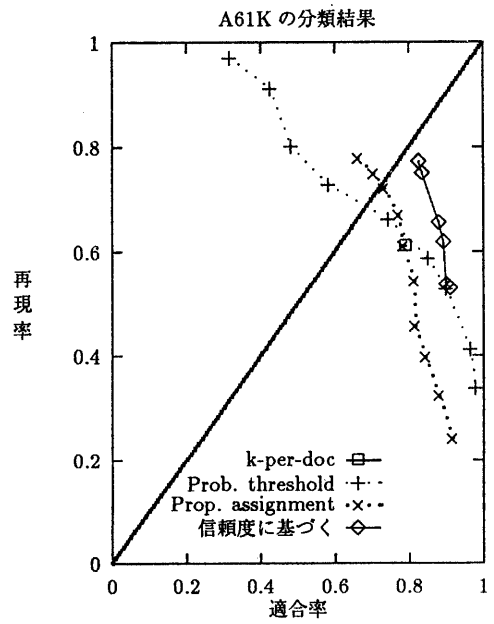

図 4: 適合率－再現率曲線（サブクラスからメイングループへの分類）

参考文献

[1] 亀田弘之, 藤崎博也：テーマ・キー概念・キーワード間の階層構造を利用する新聞記事情報の分類・検索システム, 情報処理学会論文誌, Vol. 28, No. 11, pp. 1103–1111 (1987).

[2] 細野公男, 田村俊作, 原田隆史, 諸橋正幸, 梅 田茂樹, 隅田英一郎：図書分類エキスパートシステム, 情処研報, 情報学基礎 8-4 (1988).

[3] Phillip J. Hayes, I. B. N., Peggy M. Andersen and Linda M. Schmandt (Carnegie Group, I.: TCS: AShell for Content-Based Text Categorization, in *Proc. The Sixth Conference on Artificial Intelligence Applications*, pp. 320–326 (1990).

[4] 河合敦夫：意味属性の学習結果にもとづく文書自動分類方式, 情報処理学会論文誌, Vol. 33, No. 9, pp. 1114–1122 (1992).

[5] Brij Masand, D. S. T. M. C., Gordon Linoff: Classifying News Stories using Memory Based Reasoning, in *Proc. of 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 59–65 (1992).

[6] Lewis, D. D.: An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task, in *Proc. of 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 37–50 (1992).

[7] 湯浅夏樹, 上田徹, 外川文雄：大量の文書データから自動抽出した名詞間共起関係による文書の自動分類, 情処研報, NL98-11, pp. 81–88 (1993).

[8] 徳永健伸, 岩山真：重み付き IDF を用いた文書の自動分類について, 情処研報, NL100-5, pp. 33–40 (1994).

[9] 岩山真, 徳永健伸：自動文書分類のための新しい確率モデル, 情処研報, FI33-9, pp. 47–52 (1994).

[10] 特許庁（編）：IPC 特許・実用新案 国際特許分類表, 社団法人発明協会, 第五版 (1990).

[11] 川島順, 清水美都子：国際特許分類 (IPC), 情報の科学と技術, Vol. 39, No. 11, pp. 503–510 (1989).