The 25th Symposium on Information Theory
and Its Applications (SITA2002)
Ikaho, Gunma, Japan, Dec. 10 - 13, 2002

文書自動分類における因子分析とクラスタリング手法を併用した単語選択法

Words Selection Method using Factor Analysis and Clustering for Automatic Document Classification

†斎藤 剛 Tsuyoshi SAITO　‡鈴木 誠 Makoto SUZUKI　†平澤茂一 Shigeichi HIRASAWA

Abstract—This paper discusses word selection methods for high accurate automatic text classification systems. In this paper, we shall propose a new word choosing method using factor analysis and clustering. We compare our method with a conventional technique, and we can obtain a high accurate classification system.

Keywords — Factor Analysis, Clustering, Document Classification

1．はじめに

近年，World Wide Web(WWW)の世界的普及に伴い，取得の可能な情報量が爆発的に増加している．また，ネットワーク高速化とストレージの大容量化，低価格化によって大量のデータを蓄積する環境が整いつつある．その中でテキストデータベース(Text database)と呼ばれる新しい形のデータベースの利用が，ここ数年，急速に進んでいる．従来からある文献データベースや，遺伝子情報データベース，電子化辞書に加えて，ネットワーク上に蓄積されたウェブページや，ＳＧＭＬ／ＸＭＬアーカイブ，ファイルシステム上のビジネス文書や電子メールなどの集積は，すべて広い意味でテキストデータベースである．

IT業界では，1年に数ペタバイトの割合でデータ量が増加すると見られている．企業各社には，1ペタバイトすなわち1000 テラバイトのストレージ容量が必要になってくる．このように人間の処理能力をはるかに超えたテキストデータベースの構築が可能であるにもかかわらず，個人が必要とする情報を抽出する技術は不完全である．そこで，テキストデータベースからユーザの必要に応じて情報を抽出する情報処理技術，未知の新しい知識を創出するテキストマイニングが研究されるようになってきている．テキストデータの情報処理技術には形態素解析や情報検索といった様々な研究が古くから行われてきているが，本稿ではテキストの自動分類を取り上げる．

テキスト自動分類システム構築のためには，文書を内容の類似性に基づき高い精度で分類するための索引語選択法が必要となる．しかし，一般に行われている相互情報量を用いる手法ではノイズが大きく，十分な精度を出すために多量の単語が必要となる．

そこで，本稿では同種の形態素を同一視してノイズを削減する因子分析をクラスタリングと併用した索引語選択法を提案する．提案手法を実データに適用した結果，従来手法と比較して，精度の高い文書の自動分類が可能であることが明らかになった．

†〒169-8555 東京都新宿区大久保 3-4-1 早稲田大学大学院理工学研究科 経営システム工学専攻．Dept. of Industrial and Management System Engineering, School of Science and Engineering, Waseda University, Okubo 3-4-1, Shinjuku-ku, Tokyo, 169-8555 JAPAN.

‡〒251-8511 神奈川県藤沢市辻堂西海岸 1-1-25 1410-3 湘南工科大学 情報工学科．Dept. of Information Science, Shonan Institute of Technology, Tujido Nisikaigan 1-1-25 Fujisawa-shi, Kanagawa-ken, 251-8511 JAPAN.

2．従来研究

2．1　テキスト自動分類

テキストの自動分類[1]とは，ユーザが未読のテキストを，あらかじめ決められた数種類のカテゴリに分類する，あるいはテキストにカテゴリを付与することをいう．例えば，ある新聞記事が「政治」についての記事なのか「経済」についての記事なのかを自動的に判別するシステムなどが考えられる．自動分類の基本的な手続きは以下のようになる．

1. 全テキストデータに対して形態素解析を行う．
2. 形態素解析結果に対して接辞処理を行う．
3. 形態素に索引語としての重要度ランクを付ける．
4. 入力テキストの内部表現化する．
5. テキストと各カテゴリの間の類似度を計算する．
6. テキストに最も類似したカテゴリを付与する．

2．1．1　形態素解析と接辞処理

形態素解析とは，文章から形態素と呼ばれる文章を構成する最も小さな要素を抽出し，その品詞を特定する処理をいう[1]．この処理は，英語のように単語の区切りが明確な言語の場合には難しいことはないが，日本語のように単語の区切れが明確ではない膠着言語では困難な処理となることが多い．

また，形態素の文法上の揺らぎを正規化する処理を接辞処理という．例えば，動詞，形容詞などを全て終止形に変換する，英語ならば"brothers"といった複数形の単語を単数形に変換する，といった処理である．これについては日本語，英語ともに必要となる．

代表的な日本語の形態素解析器には茶筅やＪＵＭＡＮなどが存在する．形態素解析器とは一般に接辞処理も同時に行うものを指すことが多い．

2．1．2　重要度ランク付け

一般にテキストの分類には索引語と呼ばれる単語をキーとして用いることが多い．しかし，上記の形態素解析結果をそのまま全て索引語として用いては，不要な語が多すぎる．例えば，「私」，「ある」など一般的過ぎるもの，「膠着言語」，「形態素」など希少すぎるものは共に情報量が少なく索引語としては不適である．また，品詞によるフィルタリングなどの不要語処理も行う．つまりテキストの分類精度を向上させるには索引語の選択が問題となってくるということである．そこで，単語ごとに文書集合との相互情報量を計算し値の高い順に索引語とする手法[2]が存在する．このときの索引語の数はユーザによって決定される．

[定義２．１] 相互情報量：文書集合Dと単語tの相互情報量$W(t;D)$を次式(1)で与える[3].

$$W(t;D)=\frac{TF(t)}{\sum_t TF(t)}\times\log(M/df(t)) \qquad (1)$$

$TF(t)$：全文書中の単語tの出現回数

M：学習データに含まれる文書の総数

$df(t)$：学習データの中で単語tを含む文書数 □

２．１．３ 内部表現化

図１で表される索引語文書行列Aのように学習データを内部表現化する．このとき各列は文書集合中の各文書に，各行は２．１．３項で抽出された索引語に対応する．行列要素a_{ij}は索引語t_iの文書d_jにおけるtf･idf値を表す．

$$A=\begin{array}{c} \\ t_1\\ t_2\\ t_3\\ t_4\\ t_5\\ t_6\end{array}\begin{array}{c}\begin{array}{ccccc} d_1 & d_2 & d_3 & d_4 & d_5\end{array}\\ \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15}\\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25}\\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35}\\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45}\\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55}\\ a_{61} & a_{62} & a_{63} & a_{64} & a_{65}\end{bmatrix}\end{array}$$

図１：索引語文書行列

[定義２．２] tf･idf値：文書dと単語tのtf･idf値$tf\cdot idf(t,d)$を次式(2)で与える[1].

$$tf\cdot idf(t,d)=\frac{tf(t,d)}{\sum_{s\in d}tf(s,d)}\times\left(\log\frac{M}{df(t)}+1\right) \qquad (2)$$

$tf(t,d)$：文書d中の単語tの出現回数 □

索引語文書行列の各行はその索引語の文書にわたる分布を表し，各列はその文書内の索引語の分布を表している．

この後，あらかじめカテゴリが付与されたテキスト集合を訓練データとして用いて，カテゴリごとに列ベクトルの平均をとり，各列がカテゴリ特徴ベクトルとなる索引語カテゴリ行列B(図2)を生成する．

$$B=\begin{array}{c} \\ t_1\\ t_2\\ t_3\\ t_4\\ t_5\\ t_6\end{array}\begin{array}{c}\begin{array}{ccc} c_1 & c_2 & c_3\end{array}\\ \begin{bmatrix} a'_{11} & a'_{12} & a'_{13}\\ a'_{21} & a'_{22} & a'_{23}\\ a'_{31} & a'_{32} & a'_{33}\\ a'_{41} & a'_{42} & a'_{43}\\ a'_{51} & a'_{52} & a'_{53}\\ a_{61} & a_{62} & a_{63}\end{bmatrix}\end{array}$$

図２：索引語カテゴリ行列

２．１．４ 類似度計算

図２の各列をベクトルとみなすと，各ベクトルは各カテゴリの特徴をあらわしていると考えられる．また，同様に未分類文書も特徴をあらわすベクトルをもつ．そこで，ベクトル空間モデルではテキストとカテゴリを索引語の重みベクトルで表現し，その間の類似度を計算する．各カテゴリ特徴ベクトルと未分類文書のベクトルとの類似度を評価することによって，未分類文書が格納されるべきカテゴリを計算することが可能となる．

このときの文書d_xとd_yの類似度$\sigma(d_x,d_y)$は文書ベクトル$\mathbf{x}=(x_1,x_2,\cdots,x_T)$, $\mathbf{y}=(y_1,y_2,\cdots,y_T)$の余弦(式(3))で計算される[1].

$$\sigma(d_x,d_y)=\frac{\sum_{i=1}^{T}x_i\cdot y_i}{\sqrt{\sum_{i=1}^{T}x_i^2\times\sum_{i=1}^{T}y_i^2}} \qquad (3)$$

２．２ 問題点

上記で示したように，文書の分類は文書内に出現した形態素をほぼそのまま用いて行われる．しかし，２．１．２項のように相互情報量をもって索引語を抽出した場合，カテゴリを特定する情報が極めて近いと考えられる複数の形態素を同時に索引語として採用する可能性が高い．例えば，「サッカー」という形態素と「オフサイド」という形態素は共にスポーツカテゴリを特徴付ける索引語であるが，それぞれが単独で出現することは少なく，統合してしまったほうが良いことが想像される．つまり，相関の高い複数の形態素が索引語群に入っていることで情報量が減少してしまっていると考えられる．

３．提案手法

上記の問題解決のため，因子分析を行うことで相関の高い形態素を抽出，統合し索引語群の情報量を増加させる．また，単純に因子分析を行った場合，統合されるべきではない形態素が統合されることがあるため，クラスタリングを併用する．以下でそれぞれについて簡単に説明する．

３．１ 因子分析モデル

統計の一手法に因子分析がある．因子分析とは，変数の間の相関関係に基づいて，変数群の背後に潜むいくつかの共通原因を抽出し，因子として変数群を縮約する手法である[4]. 因子分析を用いて，形態素・文書・因子の関係を定式化すると次式(4)のようになる．

$$a_{ij}=g_{i1}p_{1j}+g_{i2}p_{2j}+\cdots+g_{in}p_{nj}+v_{ij} \qquad (4)$$

a_{ij}：特定の変数．本論では２．１．３項のa_{ij}と同意

g_{im}：共通因子負荷量．本論では形態素t_iと因子f_mとの相関

p_{mj}：共通因子得点．本論では因子f_mにおける文書d_jの得点

v_{ij}：独自因子得点．

n：因子数

つまり，因子負荷量g_{im}がユーザの与える閾値以上に高い形態素t_i同士は相関が高い形態素と考え，因子f_mにまとめることが可能となる．

３．２ テキスト空間分割

通常の統計処理と異なり，テキストデータでは一変数（本論では形態素）が複数の意味を持つことがある．そのため上記の因子分析の手法のみでは，一つの形態素が複数の意味を持ちやすいバリエーション豊かな文書集合では適切な因子を求められない可能性がある．

例えば，スポーツカテゴリでは「清原」と「和博」は統合できるが，他のカテゴリにおける「清原」が「和博」と相関が高く統合するかは不定である．また，「核」という単語は，政治カテゴリでは「北朝鮮」と統合するが，産業カテゴリでは「発電」と統合するといった事例が存在する可能性がある．

つまり，同一の単語であっても文書によっては異なった因子に依存している可能性がある．その問題を解決するため，テキスト空間を分割し，それぞれのクラスタにおいて因子分析を行う方法を提案する．

３．２．１ 空間分割法

テキスト空間の分割にはクラスタリングアルゴリズム[5]を用いる．クラスタリングアルゴリズムには凝集法（ツリークラスタリング），Two-way 法（ブロッククラスタリング），k-means 法などがある[5]. 凝集法とは，要素を逐次的に大き

くなるクラスタに一緒に結びつけていく手法である．Two-way 法とは，変数とデータを同時にクラスタ化する手法である．k-means 法は k 個のクラスタを作成したいときに用いる．

k-means 法は，ランダムに生成した k 個のクラスタの中心値と各データとの距離を計算し，最も近いクラスタにデータを配分する．次に，クラスタに配分されたデータから，そのクラスタの重心値を計算する．クラスタの重心値は，そのクラスタに属する全ての点の平均値とする．これを，全てのデータに対して，データとデータが属するクラスタの重心値との距離の合計が最小になるまで繰り返す手法である．

本提案ではこの k-means 法を用いている．なぜなら，本手法におけるクラスタリングの目的は，テキスト空間をひとつの形態素がひとつの意味のみを持つようにテキスト空間を分割することにあるからである．つまり，その最適な分割数はカテゴリの数によってユーザが推定できると考えるため，分割数をユーザが決定できるアルゴリズムが適していると考えられる．

3．2．2 分割空間での形態素統合

空間分割後，それぞれの空間において因子分析を行うわけであるが空間により統合する形態素が異なるのは自明である．そこで，本提案ではある空間で統合された形態素同士が，他空間では異なった因子に閾値以上の負荷量を持つ場合，統合されないこととした．つまり，ある形態素同士が統合されるのは，他空間においても統合されている場合，もしくは他空間では双方とも相関の高い形態素を持たず統合しなかった場合に限ることとする．

3．3 テキスト空間結合

分割数をユーザが決定できる k-means 法であるが，クラスタあたりの文書数に下限は無いため，必要以上に細分化しすぎてしまうリスクがある．つまり文書数が極端に少ない場合，マクロ的にはまったく意味の無い相関が発生する可能性がある．

そこで，分割した空間のうち任意の二つを結合した空間においても同様に因子分析を行い，結合前後で形態素の統合数の変化をみる．通常，クラスタは少ないほど形態素は統合しやすくなり，空間結合後に統合数は増加する．しかし，結合前のクラスタで意味の無い相関が多数発生していた場合，結合後の統合数が減少する．よって，結合後の統合数が減少した場合，その二つのクラスタは結合するものとする．

3．4 提案アルゴリズム

形態素群を再構築し，索引語を改善するため，2．1節におけるアルゴリズムの Step2 と Step3 の間に次の処理を加える．

1. 形態素解析結果から形態素文書行列を生成する．
2. k-mean 法により k 個のクラスタに文書集合を分割する．
3. 各クラスタに対して因子分析を行い，統合できる単語数をチェックする．
4. 任意の二つのクラスタを結合してできるクラスタにおいて因子分析を行い統合できる単語数をチェックする．
5. Step4 で最も統合数が減少した二つのクラスタを結合する．
6. Step4 と Step5 を繰り返し統合数の減少がなくなった時点で Step7 へ進む．
7. あるクラスタで相関があり，任意の他クラスタでは異なった因子に閾値以上の負荷を持たない形態素を統合する．

[定義3．1] 形態素文書行列：索引語文書行列の索引語を形態素すべてと置き換えたもの．

4．シミュレーション

文書の自動分類シミュレーションをもって提案アルゴリズムの有効性を示す．

4．1 シミュレーション条件と評価基準

分類結果が既知の毎日新聞 1994 年分[6]を用いて分類精度比較を行った．上記データに形態素解析[7]を行った結果から 5 カテゴリ 1000 記事を学習データ，500 記事をテストデータとして抜き出したデータを利用した．それらに対し，本手法を用いて再構築した形態素群と，従来手法により構築した形態素群から相互情報量に基づいて索引語を抽出した．それぞれテストデータの分類を行い，F 尺度および分類精度の比較を行った．

[定義4．1] F 値：F 値とは，次式(5)で定義されるように精度と再現率を共に考慮した尺度である[1]．

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} \quad (R: \text{再現率}, \ P: \text{精度}) \qquad (5)$$

α は再現率と精度のどちらを重要視するかのパラメタで，今回は $\alpha = 0.5$ としている(つまり再現率と精度の重要性を等しく評価している)．

[定義4．2] 再現率と精度：再現率 R とは，あるカテゴリ文書中，正解のカテゴリに入った割合であり次式(6)で表される[1]．

$$R = \frac{\text{正しくカテゴリ } C \text{ に分類された文書数}}{\text{真のカテゴリが } C \text{ である文書数}} \qquad (6)$$

また，精度 P とは，あるカテゴリに入った文書中，そのカテゴリが正解だった文書の割合であり，次式(7)で表される[1]．

$$P = \frac{\text{正しくカテゴリ } C \text{ に分類された文書数}}{\text{カテゴリ } C \text{ に分類された文書数}} \qquad (7)$$

4．2 結果

(実験1) 分類精度：テストデータ 500 記事について，索引語数を 20 から 1280 まで増加させた時の分類精度（図3）及び各カテゴリでの F 値（表1）の変化を記録した．このとき図表の basic は従来手法を表しており，factanal は因子分析のみを用いた提案手法を，classl5 はクラスタ数を 5 とし因子分析を行った提案手法を表している．

図3の横軸は分類に用いた索引語数であり，縦軸は分類精度を表している．その結果，クラスタリングを行わずに因子分析を行った手法は，従来手法に対して索引語数 80 のケースを除いて精度が改悪された．しかし，クラスタリングと因子分析を併用した手法では常に精度が勝る結果を示した．

(実験2) F 値：表1では各カテゴリの索引語数に対しての F 値の変化を表している．この結果も分類精度と同様に，平均的にはクラスタリングを併用した手法が最も F 値が良く，因子分析のみを用いた手法の F 値が悪かった．

また，図4は索引語数を変化させたときのF値の平均を，手法ごと，カテゴリごとにグラフ化したものである．ここから，因子分析のみを用いた手法はカテゴリDでは高い値を出しているが，他のカテゴリでは低いことが確認された．

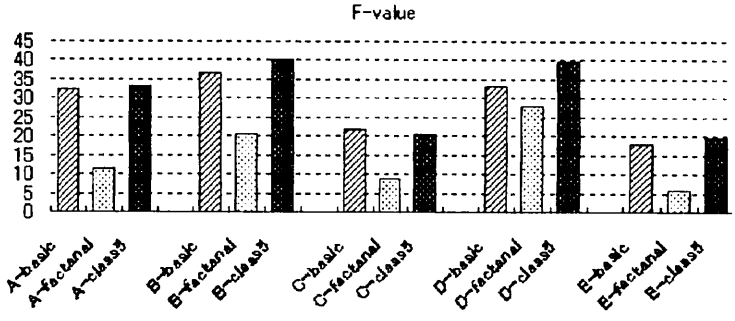

図4：カテゴリごとのF値の平均

5．考察

（考察1）因子分析のみの手法：この手法では分類精度に芳しい結果を出すことはできなかった．この原因をF値から想定してみると，因子分析の結果，統合された形態素が特定のカテゴリ，この例ではカテゴリD（スポーツ）に寄与している割合が大き過ぎたのではと考えられる．事実，統合された形態素を検証してみるとスポーツがらみのものが多かった．このことから，因子分析に

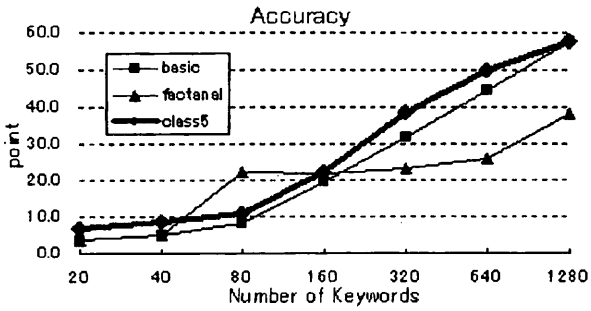

図3：分類精度

は表出しやすい単語と，しにくい単語がありクラスタリングなどの作業は必須であると考えられる．

（考察2）クラスタリングを併用した手法：この手法では索引語数が少ないとき従来手法に勝っている．つまり，ノイズが減少し効率的な索引語が抽出できたと考えられる．しかし，索引語数が1280に達した時点で従来手法に追いつかれている．これは，索引語が少ないときは形態素の統合が効いており索引語の情報量が大きくなったことから精度が改善されたが，索引語が多くなるにつれ統合せずとも充分な情報量が索引語に存在するためだと考えられる．

（考察3）計算量について：従来手法では用いない処理を加えている関係上，今回の実験で因子分析の時間が一回あたり数十秒余計に必要となる．また，クラスタリングと併用する手法では因子分析が ${}_kC_2 + k$ 回から $k^2/2$ 回必要となる．

6．結び

現在，クラスタ数の決定は経験に基づいている．この後は理論的背景に裏打ちされたクラスタ数の決定法を考えてみたい．また，クラスタリングを併用した手法には計算時間がかかるため，因子分析のみでよりよい単語集団を生成する方法を見つけたい．

謝　辞

著者の一人斎藤は，研究を進めるにあたりご討論，ご助言をいただいた早稲田大学大学院理工学研究科 大島敬志氏，同 加藤大樹氏に感謝する．

参考文献

[1] 徳永健信：“言語と計算−5 情報検索と言語処理”，東京大学出版会．
[2] Church, K.W. and Hanks, P: “Word Association Norms, Mutual Information and Lexicography”, *Proc.27th Annual Meeting of the Association for Computational Linguistics* (ACL '98), pp.76-83, 1989.
[3] 相澤彰子：“語と文書の共起に基づく「特徴量」の定義と応用”，情報学基礎自然言語処理，136-4, pp.25-32, 2000.
[4] 水野欽司：“多変量データ解析講義”，朝倉書店，1996.
[5] 宮本定明：“クラスター分析入門”，森北出版，1999.
[6] CD-毎日新聞'94，毎日新聞社，日外アソシエーツ．
[7] RWC Database RWC-DB-TEXT-96-1, RWC ワークショップ．

表1：各カテゴリのF値

F値	Keyword	20	40	80	160	320	640	1280
CategoriA	basic	8	8	10	25	37	62	75
	factanal	8	8	0	0	2	13	49
	class-factanal	8	8	10	24	47	61	74
CategoriB	basic	9	18	21	40	53	57	59
	factanal	9	18	13	10	18	26	51
	class-factanal	11	16	20	51	56	60	68
CategoriC	basic	8	8	7	16	16	38	60
	factanal	8	8	6	4	4	7	26
	class-factanal	0	0	9	14	24	42	54
CategoriD	basic	6	6	21	44	50	49	58
	factanal	6	6	37	35	35	36	42
	class-factanal	28	27	33	40	46	50	55
CategoriE	basic	6	8	13	24	28	24	23
	factanal	6	8	8	2	6	6	6
	class-factanal	6	10	15	27	17	39	26
means	basic	7.3	9.3	14.4	29.9	36.6	46.0	55.2
	factanal	7.3	9.3	12.7	10.0	13.0	17.6	34.5
	class-factanal	10.5	12.1	17.1	31.1	38.1	50.3	55.5