情 報 学 基 礎　65－4
（２００１．１１．６）

テキストの自動分類の要素分析的アプローチ

石田　栄美
emi@slis.keio.ac.jp
慶応義塾大学文学部

抄録

テキストの自動分類のメカニズムを明らかにするために、自動分類を構成する全ての要素に着目し、各要素が分類先決定にどのような影響を与えるかを検証した。まず、テキストの自動分類を構成する要素を明らかにした。要素とは、自動分類システムにおいて分類先決定に影響を及ぼす可能性がある処理のことである。次に、これら全ての要素において複数の手法を用いて、その全ての組み合わせによる分類実験を行い、各要素が分類先決定に与える影響や要素間の関係を分析した。その結果、手法を変えたことによる影響がある要素は、テキスト構造、分類先決定、単語の選択であった。要素間で交互作用がある要素は、テキスト構造と単語の選択、カテゴリ表現と分類先決定、テキスト構造と分類先決定であった。

Element analytical approach for automated text categorization

Emi　ISHIDA
Keio University

Abstract

The purpose of this paper is to clarify organization of automated text categorization. Firstly, 9 elements consisting automated text categorization are found. This paper focuses on the configuration among these elements. I prepared 5,010 Mainichi news stories for Corpus. 512 experiments are conduced for this corpus. As results of the two-way repeated-measures ANOVA, the interaction shows following elements; (1) text structure and feature selection, (2)category expression and thresholding on rank of candidate categories, (3)text structure and thresholding on rank of candidate categories.

1. はじめに

テキストの自動分類には、大きく分けて、カテゴライゼーションとクラスタリングの 2 種類の方法があるが、本研究ではカテゴライゼーションを対象とする。テキストの自動分類（text categorization）は、「テキストをあらかじめ決められた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」[1)]ことと定義できる。

テキストの自動分類に関する研究は、1960 年代[2)]から始まっているが、1980 年代まであまり

盛んに行われることはなかった。しかしながら、近年、コンピュータの発達やインターネットの普及などに伴い、電子化された資料が増加にし、自動分類の必要性はますます高まっているといえる。自動分類に関する研究も 1990 年代になり盛んに行われるようになってきた。

自動分類研究は、機械学習や情報検索分野を中心に行われているが、適用されている手法のほとんどが両分野ですでに提案された手法を自動分類に適用したものに過ぎないといえる。また、研究の中心となっているのは、分類基準の作成など自動分類に必要な様々な処理の一部分に対する手法の提案である。

つまり、テキストの自動分類は、他の分野で提案された手法を単に適用した手法で行われており、自動分類の特性を考慮した手法の提案や改良が行われているとは言いがたい。これは、既存の手法を応用する分野として自動分類が捉えられており、いまだに自動分類の特性が明らかでないためといえる。

2 既往研究

テキストの自動分類に関する研究は、英語テキストを対象としたもの [3),4)]も日本語テキストを対象としたもの [5),6)]があるが、ある一つの部分における手法の提案や比較などである。特に、分類する際に基準となるもの（カテゴリ表現）に着目した研究が多いといえる。

近年になって、カテゴリ表現、単語の選択を中心とした Yang[7)]の手法の比較研究やタイトルや件名を用いて、Larson[8)]が行った 4 種類のカテゴリ表現手法、5 種類のテキスト表現、3 種類の語幹処理法の全てを組み合わせた 60 通りの手法の比較研究、書名を用いた石田 [9)] の 2 通りの学習用データの量、3 種類の単語の切り出し手法、4 種類のカテゴリ表現手法の比較研究などがある。

これらの研究は、複数の要素を含めた比較研究を行っている点では意義がある。しかしながら、分類に効果的な手法は何か、あるいは、効果的な手法の組み合わせ何かという単純な比較に留まっており、効果的であると示された手法がどうして分類精度が高かったのかという考察に関しては、「こういう理由が考えられる」という記述があるだけで、いずれも充分な検討が行われていない。そのため、比較した手法の中で有効な手法を特定することはできるが、分類のメカニズムを明らかにできてはおらず、また、自動分類に関わる処理全てを視野に入れているとも言いがたい。

3 要素分析的アプローチ

3.1 要素分析的アプローチ

本研究は、テキストの自動分類のメカニズムを明らかにするために、テキストの自動分類を構成する要素に着目し、各要素が分類先決定にどのような影響を与えるかを検証する。

そのために、まずテキストの自動分類を構成する要素を明らかにした。要素とは、自動分類システムにおいて分類先決定に影響を及ぼす可能性がある処理のことである。次に、これら全ての要素を用いて、各要素において複数の手法を用いて、その全ての組み合わせによる分類実験を行った。各要素において手法を変えた分類実験の結果を分析することにより、分類先決定に影響を与えている要素や各要素間の関係が明らかになると考えている。

このように、自動分類全体を視野に入れた分類実験を行うことにより、分類先決定に要素が与える影響、要素間の関係を分析することを、要素分析的アプローチとする。

以下では、まず、自動分類を構成する要素をあげ、その要素を用いた分類実験について述べる。

3.2 自動分類を構成する要素

テキストの自動分類は、シソーラスや辞書など外部の知識を用いる場合もあるが、一般的には、分類済みのテキスト集合から分類の際に基準となるものを作成し、それをもとに分類対象テキストを作成するという手順で行われる。分類済みのテキスト集合を用いた自動分類を行う場合には、ほとんどの場合、この手順を踏まなければならない。

既往研究の調査や実際の分類実験から、自動分類を構成する要素を洗い出した [10)]。自動分類は、大きく分けると、学習フェーズ、分類フェー

ズ、評価フェーズの 3 つにわけることができる。学習フェーズは、分類済みのテキスト集合から分類する際に基準となるもの（カテゴリ表現）を作成する部分であり、分類フェーズは実際に分類対象テキストを分類する部分である。また、既往研究では評価フェーズを自動分類システムの中に含めている例は少ないが、自動分類全体を視野に入れる場合には評価フェーズも必要であると考え含めることとした。

各フェーズおける要素の一覧を図 1 に示す。以下では、各フェーズにおける要素について説明する。

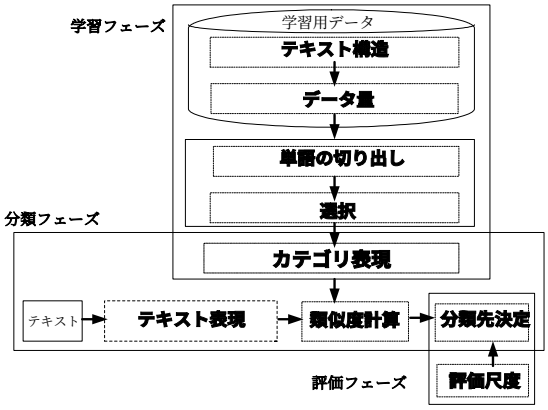

図1　各フェーズにおける要素

3.2.1　学習フェーズ

学習フェーズでは、分類先を決定する際に基準となるカテゴリ表現を作成する。分類する際の基準となるものは、一般的に分類基準、分類ルール、辞書などと呼ばれることがあるが、本研究ではカテゴリ表現と呼ぶ。カテゴリ表現は、カテゴリに分類済みのテキスト集合から、単語と各カテゴリがどの程度関連しているかを表現したものである。

分類済みのテキスト集合を、学習用データと呼ぶ。学習用データにおいて要素となるのは、テキスト構造とデータ量である。

テキスト構造では、テキスト中のどの部分を用いるかということで手法が異なるといえる。例えば、新聞の見出しだけを用いて分類を行う場合と、リード、または全文を用いる場合でどのような違いがあるかということである [11)]。テキストのどの部分を用いるかによっても分類精度が異なる可能性があるので、一つの要素として考えられる。

カテゴリ表現を行うときに用いるデータの量によってより適切な表現が行える可能性があるので、データ量も一つの要素である。

カテゴリ表現はテキストから得られた単語集合を用いて表現するが、日本語テキストの場合は連続した文字列で表現されているので、文字列から単語を切り出すことが必要となる。単語は分類の手がかりとなる最少単位なので、切り出し手法も分類精度に影響する可能性がある。英語テキストでは語幹処理を行う場合がある。

次に、得られた単語集合から単語の選択を行う場合がある。これは、カテゴリ表現がより適切になるようにノイズとなる単語を削除するためにや計算量を減らすために行われる。

以上のように学習フェーズでは、テキスト構造、データ量、単語の切り出し・語幹処理、単語の選択、カテゴリ表現の 5 つの要素がある。

3.2.2　分類フェーズ

分類フェーズでは、実際にテキストをカテゴリに分類する。分類対象テキストは、テキスト中の単語とカテゴリ表現を照合し、類似した特徴を持つカテゴリに分類される。そのためには、分類対象テキストをカテゴリと照合できるように変換することが必要であり、この処理をテキスト表現と呼ぶ。テキスト表現は、分類対象テキスト中で出現する単語の集合であり、単語の切り出しや単語の選択が行われ、単語に出現回数や重み情報が付与されたりする。これらの処理を含めて、テキスト表現とする。

分類対象テキストと類似した特徴をもつカテゴリを求めるためには、カテゴリ表現とテキスト表現の類似度を計算して求める方法がある。この類似度計算方法によって、類似するカテゴリが異なる場合があるので、類似度計算も一つの要素となる。

また、分類先の決定方式は、テキストがあるカテゴリに分類されるか、されないかを分類結果として示す 2 値分類方式とテキストが分類されるべきカテゴリ順にランキングするランキング分類方式がある。これも分類先決定に大きく関わるので要素となる。

分類フェーズでは、テキスト表現、類似度計算、

分類先決定の 3 つの要素がある。

3.2.3 評価フェーズ

評価フェーズは、システムが分類先を決定する方法に従い、評価方法が異なる。2 値分類の評価方法は、再現率、精度、フォールアウト、成功率、エラー値などがある。ランキング方式には、精度、再現率、11 ポイント平均精度などがある。この評価尺度もひとつのフェーズとなる。

4 分類実験

4.1 実験の概要

各要素で提案されている手法を用いて、それらの組み合わせによる分類実験を行った [12]。各要素において異なる手法を用いるが、その手法のうちどの手法が分類に有効であるかを調べるのではなく、各要素において用いる手法を変えることによって分類精度に違いが表れるかを調べることを目的としている。

実験では自動分類全体の流れに眼においている。各要素において提案されている全ての手法の全ての組み合わせによる実験を行うことは困難である。

ここでは、各フェーズの全要素である 9 要素のうち、7 要素（類似度計算と評価尺度を除く）において提案されている代表的な手法を 2 種類用い、それらの全ての組み合わせにおける分類実験を行った。実験は全部で 512 通りである。

以下では、実験に用いたテストコレクション、用いた手法について述べる。

4.2 日本語新聞記事テストコレクションの作成

欧米での自動分類研究では、英語の新聞記事のテストコレクションである Reuters-21578 が最も用いられているといえる。このテストコレクションは、ロイター通信の記事 21,578 件からなり、分類に用いられてきた Topics というカテゴリセットには 135 のカテゴリがある。

このように英語テキストのテストコレクションはあるが、日本語のテキストには分類実験に用いることのできるコレクションは少ない。

本研究では、分類実験に適当な日本語のテストコレクションを作成した。テキストには新聞記事を用い、各記事に分類カテゴリを割り当てた。

「毎日新聞 CD-ROM データ集」の 1994 年と 1999 年版を用いた。これは、毎日新聞東京・大阪本社発行の各 1 年分の記事を収録したものである。このうち 1994 年の 6 月分と 1999 年の 10 月分を対象に分類カテゴリを付与した。

分類カテゴリは、毎日新聞縮刷版の記事索引で用いられている分類カテゴリをそのまま用いた。分類カテゴリは第一階層から第三階層まであり、第一階層のカテゴリは、政治、外交、経済、労働、社会など 10 カテゴリであり、第三階層までを含めた総カテゴリ数は 309 である。

新聞記事に分類カテゴリを割り当てる作業では、記事索引中の見出しをもとに、それに該当する記事をデータ集から検索し、同定できるものだけに分類カテゴリを付与した。

本研究では、このうち 1994 年の 6 月分を用いて分類実験行う。分類カテゴリが付与できたものは 5,010 件であり、4,008 件を学習用データに、残りの 1,002 件を評価用データに用いた。記事には見出し、リード、本文などのタグが付与されているので、各タグにおける平均文字数などを表 1 に示す。

表1 基本的なデータ

	件数	見出しの平均文字数	リードの平均文字数	本文の行数	本文中(1行)の平均文字数
全データ	5,010	26.6	110.1	6.0	76.2
学習用	4,008	26.8	111.2	5.8	80.3
評価用	1,002	26.6	110.3	6.0	77.0

4.3 各要素で用いた手法

各要素内で用いた手法を図 2 に示す。以下では、各手法について具体的に述べる。

テキスト構造では、学習用データで見出しのみを用いた場合と全文を用いた場合の 2 通りの方法を用いた。

学習用データ量では、4,008 件と 2,004 件の 2 通りを用いて実験した。

単語の切り出しには、形態素解析システム「茶筅」[13]と n-gram（n=2）の 2 通りを用いた。

単語の選択は、形態素解析システムの場合は切り出した単語全てを用いる場合と、その中から名詞だけを用いる場合の 2 通りで実験を行った。n-gram の場合も全ての単語を用いる場合と漢字だけで構成されている文字列だけを用いる場合の 2 通りで実験を行った。

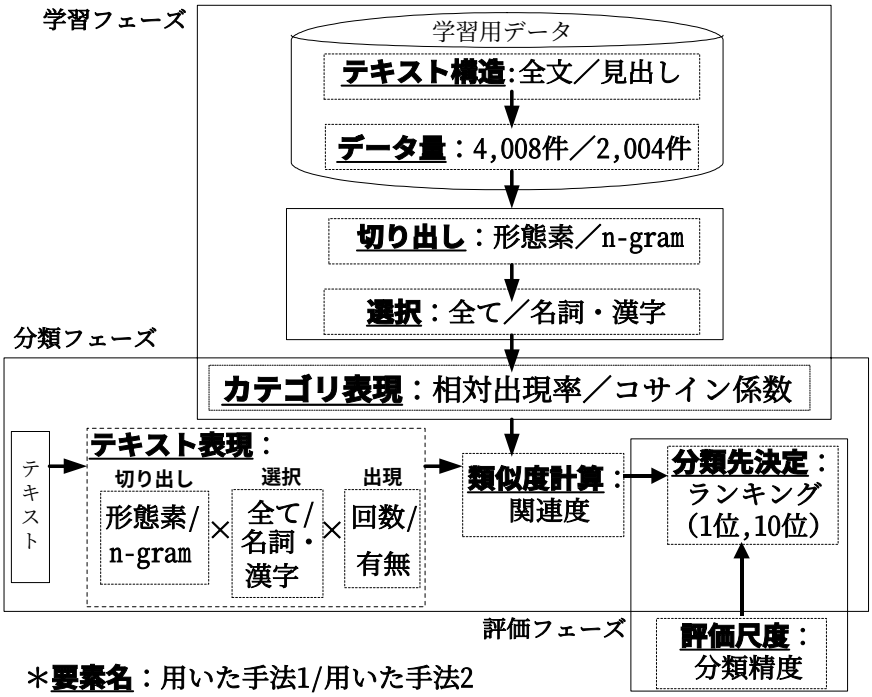

図2　各フェーズの要素内で用いた手法

カテゴリ表現は、図書に NDC カテゴリを分類する実験[9]で最も精度が高かった手法（相対出現率法）と岸田[14]が雑誌論文の表題を用いて分類記号の付与を行ったときに最も精度が高かった手法であるコサイン係数を用いた。カテゴリ表現に関しては、多数の手法が提案されているため、これらの手法以外にも比較実験を行わなければならない手法があるが、本発表では日本語テキストを用いた分類実験で他の手法との比較実験の結果、有用性を示しているという点で上の 2 つの手法を用いた。これらの手法によるカテゴリ表現は単語の各カテゴリに対する重み計算をすることによって行う。各手法のカテゴリ表現方法は以下の通りである。

(1) 相対出現率法

相対出現率法によるカテゴリ表現は、カテゴリ C_i（i= 1,2,3,・・・・,N）における単語 t_j（j= 1,2,3,・・・・,M）の重み w_{ij} は、

$$w_{ij} = \frac{T_{ij}}{\sum_{i=1}^{N} T_{ij}}$$

で求める。ここで、T_{ij} は単語 t_j のカテゴリ C_i における出現回数である。

(2)コサイン係数

カテゴリ C_i（i= 1,2,3,・・・・,N）における単語 t_j（j= 1,2,3,・・・・,M）の重み w_{ij} は、

$$w_{ij} = \frac{d_{jci}}{\sqrt{d_j d_{ci}}}$$

で求める。ここで、d_j は単語 t_j が出現するテキスト数、d_{ci} はカテゴリ C_i に属するテキスト数、d_{jci} は単語 t_j が出現するテキスト数の中でカテゴリ C_i に属するテキスト数である。

テキスト表現を行うためには、分類対象テキストに対して学習用データと同様、単語の切り出し、単語の選択が必要となる。これには、学習用データに対して用いた方法と同じ手法を適用した。また、テキスト表現中での重みは、分類対象テキスト中に出現する単語の有無（出現すれば 1、出現しなければ 0）を用いた場合と出現回数をそのまま重みにした場合で行った。

分類対象テキストとカテゴリ表現との類似度計算は、記事とそれぞれの分類カテゴリに対する重みの総和を計算することによって求める。ここでは、以下の方法を用いた。

各カテゴリ C_i の特徴ベクトルを $c_i = \{ w_{i1}, w_{i2}, \cdots w_{ij} \}$ とし、分類対象テキストの特徴ベクトルを $q_i = \{ w_{q1}, w_{q2}, \cdots w_{qj} \}$（$w_{qj}$ は分類対象テキスト中

に単語 t_j の出現回数とする。）と表わすと分類対象テキストの各カテゴリ対する関連度は、

$$関連度 = \sum_{j=1}^{M} w_{ij} w_{qj}$$

で求める。この結果として、分類対象テキストと各カテゴリの類似度がわかる。

分類先は、ランキングにより決定した。分類対象テキストと各カテゴリの類似度計算を行うことによって、テキストと類似している順にカテゴリがランク付けされる。1 位にランク付けされたものだけを正解とする場合(1 位)と 10 位までにランク付けされたものを正解とする場合(10 位まで)という 2 つの方法で分類先を決定した。

評価尺度は、実験結果のうち、すでに付与されているカテゴリと一致した割合とした。

5 実験結果の分析

5.1 実験結果

分類実験の結果の一部を表 2~5 に示す。この表は、512 通りのうち 128 通りの実験結果を示したものである。表 2 は単語の出現回数でテキスト表現を行った分類結果のうち、1 位にランク付けされたものだけを正解とした場合の分類精度（1 位）であり、表 3 は 10 位までを正解とした（10 位）場合の分類精度である。表 4、表 5 は単語の有無でテキスト表現した場合のそれぞれ 1 位、10 位の分類精度である。

この表から、各要素に着目して手法に違いによる分類精度を比較してみると、テキスト構造では全文よりも見出しを用いた場合の分類精度が高いといえるが、その他の要素では一定の傾向は見られないことがわかる。

つまり、テキスト構造に関しては有効な手法があるといえるが、その他の要素については、精度が高いものでも低いものでも同じ手法を用いている場合が多く、ある特定の手法が有効であるということはいえない。手法の組み合わせが分類精度に影響を及ぼしており、各要素で用いる手法が複雑に関係しているのではないかということがいえる。

表2 テキスト表現が出現回数の分類結果(1位)

カテゴリ表現	テキスト構造	データ量（件）	単語の切り出し（単語の選択） 形態素（全て）	形態素（名詞）	n-gram（全て）	n-gram（漢字）
相対出現率	見出し	4,008	60.1	61.9	62.5	55.5
		2,004	55.2	57.2	57.7	49.9
	全文	4,008	15.4	27.3	8.8	25.7
		2,004	13.1	25.2	8.9	22.3
コサイン係数	見出し	4,008	55.9	60.3	65.1	53.8
		2,004	53.2	57.3	59.7	47.5
	全文	4,008	6.7	30.1	11.8	29.7
		2,004	7.5	26.8	8.6	25.4

表3 テキスト表現が出現回数の分類結果(10位)

カテゴリ表現	テキスト構造	データ量（件）	単語の切り出し（単語の選択） 形態素（全て）	形態素（名詞）	n-gram（全て）	n-gram（漢字）
相対出現率	見出し	4,008	82.6	82.8	83.1	80.3
		2,004	79.4	78.3	80.1	74.9
	全文	4,008	52.1	72.0	41.7	67.3
		2,004	51.4	70.1	43.0	65.3
コサイン係数	見出し	4,008	80.4	82.2	82.7	78.7
		2,004	78.7	78.4	80.7	74.3
	全文	4,008	29.5	65.7	34.7	67.4
		2,004	29.7	64.0	33.2	64.3

表4 テキスト表現が出現の有無の分類結果(1位)

カテゴリ表現	テキスト構造	データ量（件）	単語の切り出し（単語の選択） 形態素（全て）	形態素（名詞）	n-gram（全て）	n-gram（漢字）
相対出現率	見出し	4,008	60.3	62.3	62.6	55.5
		2,004	55.5	57.4	57.4	49.9
	全文	4,008	14.8	26.2	11.3	28.0
		2,004	14.6	24.3	11.5	24.5
コサイン係数	見出し	4,008	56.0	60.7	65.0	53.4
		2,004	53.4	57.2	59.7	47.6
	全文	4,008	11.4	37.1	12.3	34.0
		2,004	10.2	33.2	10.5	29.8

表5 テキスト表現が出現の有無の分類結果(10位)

カテゴリ表現	テキスト構造	データ量（件）	単語の切り出し（単語の選択） 形態素（全て）	形態素（名詞）	n-gram（全て）	n-gram（漢字）
相対出現率	見出し	4,008	82.3	82.6	83.2	80.2
		2,004	79.3	78.4	79.8	74.9
	全文	4,008	63.3	75.3	45.6	70.5
		2,004	62.3	73.6	46.5	70.5
コサイン係数	見出し	4,008	81.1	82.3	82.7	78.6
		2,004	78.5	78.5	80.8	74.3
	全文	4,008	38.7	72.5	35.8	71.6
		2,004	36.5	69.3	35.6	68.1

表6　「テキスト構造」要素と「単語の選択」要素間の分散分析結果

変動要因	変動	自由度	分散	観測された分散比	P-値	F 境界値
テキスト構造	30269.17	1	30269.17	115.14	2.1E-19	3.92
単語の選択	2928.76	1	2928.76	11.14	1.1E-03	3.92
交互作用	4850.19	1	4850.19	18.45	3.5E-05	3.92
繰り返し誤差	32597.24	124	262.88			
合計	70645.37	127				

5.2　構成要素間の関係分析

5.2.1　分析手法の概要

次に、要素間にどのような関係があるかを統計的手法により分析した[15]。

分析手法には、分散分析のうち、繰り返しのある二元配置分散分析を用い、各要素で用いた手法の違いによる効果、2 つの要素間の関係を調べた。この分散分析手法は、2 つの要素による影響の度合を調べることができ、それぞれの要素の効果を比較したり、どの要素にも差がないという仮説を検定したりすることができるものである。

2 つの要素を対象に二元配置分散分析を行うと、表 6 に示した結果が得られる。これは、「テキスト構造」と「単語の選択」の 2 つの要素を対象に分析を行った結果である。この表において、「観測された分散比」が「F 境界値」よりも大きく、「P 値」が 0.05（有意水準を 0.05 とする。）より小さければ、手法を変えた効果や交互作用があるといえる。交互作用とは、要素のお互いの影響、組み合わせたことにより生まれる効果のことである。

5.2.2　分析結果

2 つの要素の組み合わせて分散分析を行った結果、用いた手法により影響がある要素、交互作用があるとされた要素の組み合わせを表 7 に示す。

用いた手法により影響がある要素は、「テキストの構造」、「分類先決定」と「単語の選択」であった。

また、交互作用がある要素は、「カテゴリ表現」と「分類先決定」、「テキスト構造」と「単語の選択」、「テキスト構造」と「分類先決定」であった。

6　考察

実験結果の分析から、手法を変えたことによる影響がある要素は、テキスト構造、分類先決定、単語の選択であった。テキスト構造は全体的な傾向からも明らかであり、単語の選択は統計的分析で明らかになった要素である。この結果から、自動分類において、テキストの構成単位のどの部分を用い、どのような単語の選択を行うかが分類精度に大きな影響を与えることがいえる。分類先決定は、テキストと類似したカテゴリがランキングで出力されるときにどの範囲までを正解とするかを決定する要素なので、この要素で異なる手法を用いれば分類精度に影響を及ぼすのは当然である。

「テキスト構造」と「単語の選択」、「テキスト構造」と「単語の選択」、「テキスト構造」と「分類先決定」はお互いに影響しあう要素であり、他の要素間では影響しあう関係はないことが明らかになった。

表7　分散分析の結果

要素1	要素2	要素1の手法による違い	要素2の手法による違い	交互作用
カテゴリ表現	分類先決定	×	○	○
テキスト構造	単語の選択	○	○	○
テキスト構造	分類先決定	○	○	○
カテゴリ表現	単語の切り出し	×	×	×
カテゴリ表現	単語の選択	×	○	×
カテゴリ表現	テキスト表現	×	×	×
単語の切り出し	単語の選択	×	○	×
データ量	カテゴリ表現	×	×	×
データ量	単語の切り出し	×	×	×
データ量	単語の選択	×	○	×
データ量	テキスト表現	×	×	×
データ量	分類先決定	○	○	×
テキスト表現	単語の切り出し	×	×	×
テキスト表現	単語の選択	×	○	×
テキスト表現	分類先決定	×	○	×
テキスト構造	カテゴリ表現	○	×	×
テキスト構造	単語の切り出し	○	×	×
テキスト構造	データ量	○	×	×
テキスト構造	テキスト表現	○	×	×
分類先決定	単語の切り出し	○	×	×
分類先決定	単語の選択	○	○	×

従来の自動分類研究ではある一つの要素だけに注目し、その要素の中だけで手法の提案や比較を行っている研究が多かった。本研究の実験結果の分析から、分類先決定に影響を与える要素、要素間で影響を与えあう要素が明らかになった。このことは、一つの要素だけでなく、要素間の関係も考慮した上で自動分類を捕らえなければならないことを示唆している。

7 おわりに

本研究では、自動分類のメカニズムを明らかにするために、自動分類全体を視野に入れた要素分析的アプローチを提案し、それに沿った実験を行った。その結果、自動分類を構成する要素が分類先決定に与える影響や要素間の関係を明らかにした。

しかしながら、今回の分析では、自動分類を構成する要素を対象に、どの要素が分類先決定に大きな影響を及ぼすか、要素間の関係があるかないかの 2 点を焦点にしており、要素において用いた手法が変化した場合の分類精度だけに着目している。

今後は、見出しを用いた場合、全文を用いた場合、単語の選択を行った場合、行わない場合で、単語集合の大きさや特性にどのような影響があるのかなど、手法の違いではなく、その手法が意味するデータの変化を詳細に調べることが必要である。

また、これらの実験は用いた手法にも大きく影響することが予想される。今回の実験において、カテゴリ表現で用いた手法は情報検索分野で提案された手法であり、機械学習分野で提案された手法を用いていない。この 2 つの分野で提案されている手法はアプローチが大きく異なるので、今後は、機械学習分野で用いられている手法を用いた実験を行い、さらに分析していくことが必要である。

【引用文献】

1) 徳永健伸. 情報検索と言語処理. 東京, 東京大学出版会, 1998
2) Maron, M.E. "Automatic indexing: An Experimental inquiry. " Journal of American computer Machinery. Vol.8, pp.404-417(1961)
3) Iwayama, M., Tokunaga, T. "A probabilistic model for text categorization: based on a single random variable with multiple values." In Proceedings of 4th Conference on Applied Natural Language Processing, pp.162-167(1994)
4) Apte, C., Damerau, F., Weiss, S. M. "Automated Learning of decision rules for text categorization." ACM Transaction of Information Systems, Vol.12, No.3, pp.223-251(1994)
5) 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, p.1114-1112(1992)
6) 藤井洋一 他. "共起情報を利用した文書の自動分類. " 情報処理学会自然言語処理 118-16, p.97-104(1997)
7) Yang, Y., "An Evaluation of statistical approaches to text categorization. " Journal of Information Retrieval. Vol.1, No. 1/2, pp.67-88(1999)
8) Larson, R. Ray. "Experiments in Automatic Library of Congress Classification." Journal of The American Society for Information Science. Vol.43, No.2, pp.130-148(1992)
9) 石田栄美. "図書を NDC カテゴリに分類する試み. " Library and Information Science. No.39, pp.31-45(1998)
10) 石田栄美. "テキストの自動分類を構成する要素 " 2000 年度三田図書館・情報学会研究大会発表論文集, pp.45-48(2000)
11) 石田栄美. "日本語テキストの構成単位を利用した自動分類. " 1999 年度第 47 回日本図書館情報学会研究大会発表要綱. pp.37-41(1999)
12) 石田栄美. "構成要素全体から考えるテキストの自動分類 ～日本語新聞記事テストコレクションによる分類実験～. "日本図書館情報学会 2001 年度春季研究集会発表要綱. pp.51-54(2001)
13) 形態素解析システム 茶筅 .http://chasen.aist-nara.ac.jp/index.html.ja
14) 岸田和明. “論文標題に基づく分類記号をディスクリプタの自動付与”, 日本図書館情報学会 2000 年度研究大会発表要綱, p.110-113(2000)
15) 石田栄美. "テキストの自動分類に関わる構成要素間の関係の分析 " 2001 年度三田図書館・情報学会研究大会発表論文集, pp.49-52(2001)