テキストの自動分類に関わる構成要素間の関係の分析

石田栄美（慶應義塾大学非常勤講師）
emi@slis.keio.ac.jp

1　はじめに

テキストの自動分類研究は、情報検索分野や機械学習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応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。そのため、分類精度の画期的な向上には至っていない。テキストの自動分類を研究する際には、まず自動分類の特性や仕組みなどを明らかにすることが必要である。

本研究では、自動分類の仕組みを明らかにするために、自動分類全体を視野に入れた研究を行い、その全体像を明らかにすることを目的としている。

現在までに、自動分類全体を視野に入れた研究の具体的な方法を提案し、その方法にそった分類実験を行ってきた [1),2)]。本発表では、分類実験結果の単純な比較だけでなく、統計的分析手法を用いてさらに詳細に分析を行った結果について述べる。

2　自動分類全体を視野に入れた研究

自動分類全体を視野に入れた研究とは、自動分類に関わる要素をできるだけ網羅的に挙げ、各要素がどの程度分類先決定に影響を及ぼすか、各要素間にはどのような関係があるのかを明らかにすることである。

テキストの自動分類は、すでに分類済みのテキストを用いて自動的に分類基準を作成し、その分類基準をもとに、分類対象テキストを既存のカテゴリに分類することである。これを行うためには、大きく分けて、分類基準を作成する部分と実際に分類対象テキストを分類する部分に分けることができる。本研究では、これらをそれぞれ学習フェーズ、分類フェーズと呼ぶ。さらに、自動分類全体を視野に入れた場合には分類結果を評価する評価フェーズが必要となる。テキストの自動分類は、この3つのフェーズから構成されていると考える。

各フェーズでは、様々な処理が必要となる。例えば、学習フェーズでは、(1)分類済みのテキストからどの構成単位を分類基準作成のために用いるかを決め、(2)用いるテキストの文章から単語を切り出し、(3)単語集合の中から分類基準に用いる単語を選択し、(4)分類基準を作成する、などの処理が必要となる。これらの処理を本研究では要素と呼び、各要素をそれぞれ、(1)テキストの構造、(2)単語の切り出し・語幹処理、(3)単語の選択、(4)カテゴリ表現とする。

自動分類研究の論文を検討した結果、自動分類システム内で行われている流れや要素は共通しており、要素は9つあることが明らかになった [1)]。

本研究では、自動分類の全体像をつかむために、自動分類に関わるこの9つの要素が分類結果にどのように影響しているのかを調べた。まず、各要素で提案されている手法を組み合わせた実験を行った [2)]。

この実験では、各要素において提案されている手法の中で最も有効な手法を調べる

のではなく、各要素は分類にどのような影響を与えるのか、また、手法の組み合わせがどのような結果をもたらすか、要素間で影響を与えあうものがあるかなどを検証する。一つの要素だけでなく、要素全体を視野に入れた上で要素間の関係を見ることが目的である。

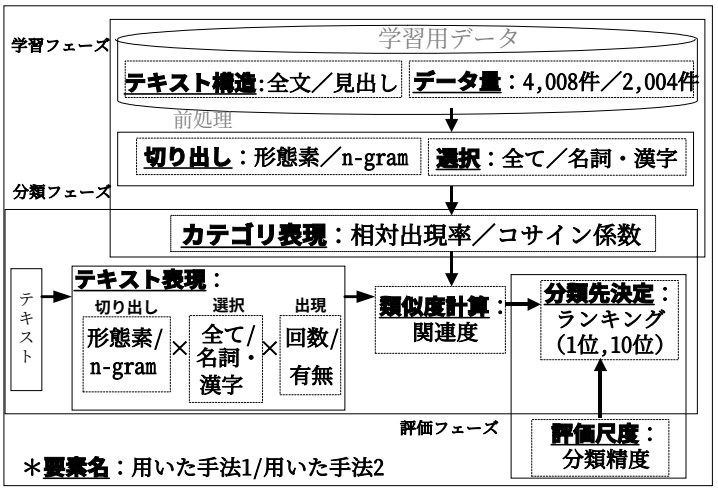

図1　各フェーズの要素と要素内で用いた手法

3　分類実験の概要

実験に用いたデータは、「毎日新聞 CD-ROM データ集」の 1994 年 6 月分の全文記事に、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（309 カテゴリ）を付与した新聞記事 5,010 件である。このうち、4,008 件を学習用データ（カテゴリ表現作成用）に、残りの 1,002 件を評価用データ（テキスト表現を行う分類対象テキスト）に用いた。

このデータを用いて、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[2)]。実験は全部で 512 通りである。

4　実験結果

分類実験の結果の一部を表 1 に示す。この表は、512 通りのうち 32 通りの実験結果を示したものである。この表から、テキスト構造に全文よりも見出しを用いた方が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ず、手法の組み合わせが分類精度に影響している。

また、512 通りの実験結果で分類精度が高かった上位 3 位と低かった下位 3 位、それぞれの手法の組み合わせを表 2 に示す。

この表を見ると、テキスト構造は、分類精度が高いときは見出しを用い、分類精度が低いときは全文を用いていることがわかる。一方、データ量に関しては、分類精度が高い方にも低い方にも、4,008 件を用いた場合と 2,004 件の場合がある。テキスト構造以外の要素に関しては、データ量同様、特定の手法を用いた場合に一定の傾向が見られない。

実験結果の単純な比較から、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。

5　構成要素間の関係分析

5.1　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。

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 3 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。

表1 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

5.2　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、交互作用があるとされた要素の組み合わせを表 4 に示す。

表から、「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」には交互作用があることがわかる。「テキスト構造」と「単語の選択」に交互作用があるとは、例えば、全文を用いたときには単語の選択を行った方が分類精度が高いことなど、2 つの要素が影響しあうこと示している。「分類先決定」要素は交互作用があるとされたが、この要素は分類精度を出すための評価尺度であり、要素間の手法を変えた場合の効果を分析する必要のない要素であるといえる。

また、各要素において手法を変えたことによる影響があるかを調べた結果、「テキストの構造」、「分類先決定」と「単語の選択」については、要素間で手法を変えることによる影響があるとされた。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。テキスト構造に関しては、結果の単純な比較や全体的な傾向 [2)] からも明らかになっている。

表2　分類精度の上位・下位3位の手法の組み合わせ

	上位			下位		
	1位	2位	3位	1位	2位	3位
分類精度	65.1	65.0	62.6	6.7	7.5	8.0
テキスト構造	見出し	見出し	見出し	全文	全文	全文
データ量	4,008件	4,008件	4,008件	4,008件	2,004件	2,004件
単語の切り出し	n-gram	n-gram	n-gram	n-gram	n-gram	形態素
単語の選択	全て	全て	全て	全て	全て	全て
カテゴリ表現	コサイン	コサイン	相対出現率	コサイン	コサイン	相対出現率
テキスト表現	出現回数	出現の有無	出現の有無	出現回数	出現回数	出現回数

表3　「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				

6　まとめ

分析結果から、「テキスト構造」と「単語の選択」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。

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比較を行っている研究が多いが、この結果から、一つの要素だけでなく、要素間の関係も考慮しなければならないことが明らかになった。

今回の分析では、要素間の関係があるかないかを焦点に統計的分析を行った。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、データ量や単語集合にどのような影響があるのかなど、データの中身に対する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) 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)

2) 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)

表4　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	○	○	×