テキストの自動分類を構成する要素

石田栄美（慶應義塾大学大学院）
emi@slis.keio.ac.jp

1　はじめに

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 年代から始まっているが、離散的な研究が多いといえる。

本研究では、テキストの自動分類に関わる要素にはどのようなものがあり、その要素に関してどのような研究が行われているかを調べることにより、自動分類研究の現状を把握し、今後の研究の課題について考察する。

まず、既往研究や実際の自動分類実験から自動分類で行われる処理の流れを追うことにより、自動分類に関わる要素を洗い出す。要素とは、自動分類の流れの中で分類先決定に影響を及ぼすと考えられる処理のことである。次に、その要素に注目して、自動分類研究を分類することにより、洗い出した要素が妥当であること、また、これからの自動分類研究における課題を示す。

2　自動分類を構成する要素

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。自動分類は、大きく分けると、3 つのフェーズにわけることができる。以下では、各フェーズにおける要素について説明する。

自動分類では、テキスト中の単語の出現傾向をもとに、テキストがどのカテゴリに属する確率が高いかを求めることにより分類先のカテゴリを決定する。学習フェーズでは、この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。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示したものである。このテキスト集合を、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。分類を行う際に、論文のタイトルを用いるか、抄録を用いるか、全文を用いるかによって、分類の精度が異なる可能性がある。そこで、テキストの構成単位の中で、どの部分を用いるのが分類にとって効果的であるのかを調べる[2)]ことが、テキスト構造にあたる。また、カテゴリ表現を充分に行うために必要なデータ数はどの程度であるのかを調べるのが、データ量[3)]にあたる。

学習用データから得られた単語集合に対しては、以下のような処理を行う場合がある。まず、英語テキストでは語幹処理を行う場合がある。

次に、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、計算量を減らすために行われる。

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、必要であれば、単語の選択が行われたり、単語に出現回数や重み情報が付与されたりする。

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[5]がある。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[1),5)]。

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。

これらの処理を示したものを図1に、各フェーズおける要素の一覧を表1に示す。

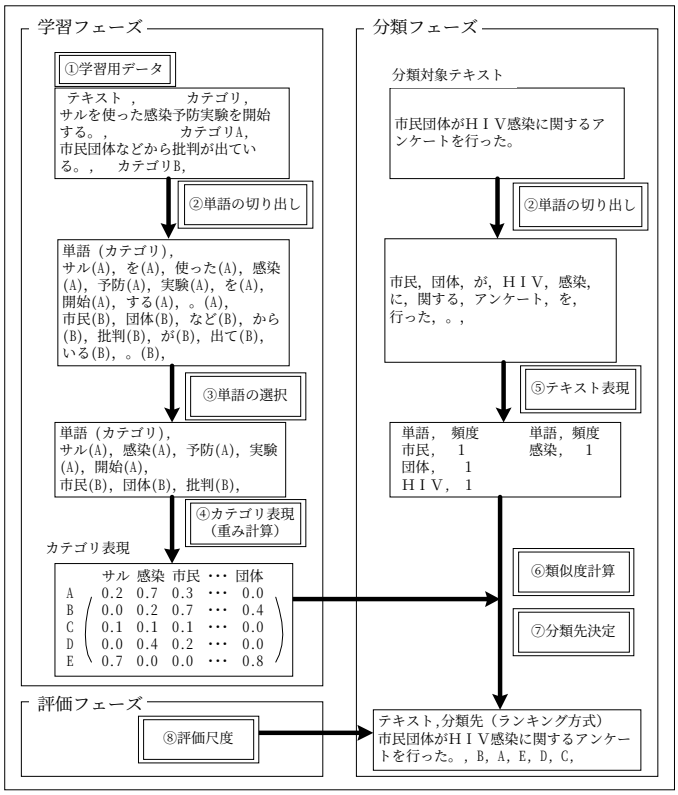

図1 ベクトルモデルにおける自動分類の手順

3　既往研究

現在までの研究では、図1の④にあたるカテゴリ表現手法の提案が中心に行われてきた。最近になって、提案された手法の中で効果的な手法を求めるための比較研究が行われるようになった。

ここでは、既往研究を、自動分類の手順という観点から、2で示した要素に着目して、1)一つの要素だけに着目し、一つの手法を提案しているもの、2)一つの要素だけに着目し、その要素に関わる複数の手法の比較をしているもの、3)複数の要素に着目し、複数の要素にわたって手法の比較をしているもの、の 3 つに分けた。

1)　一要素内、一手法の提案

現在までに、カテゴリ表現に関する様々なモデルが提案されてきている。代表的なものには、ベクトル空間モデル[6),7),8)]、確率モデル[9),10)]、ルールに基づくモデル[11)]などがある。

表1 要素の一覧

フェーズ	ステップ	要素
学習フェーズ	学習用データ	テキスト構造
		データ量
	前処理	単語の切り出し・語幹処理
		単語の選択
	カテゴリ表現	カテゴリ表現手法
分類フェーズ	テキスト表現	テキスト表現手法
	類似度計算	類似度計算方法
	分類先決定	分類先決定方法
評価フェーズ	評価尺度	評価尺度

ここでは、ベクトル空間モデルについて述べる。ベクトル空間モデルは、カテゴリとテキストをベクトルで表現し、カテゴリは一般的に単語と単語の各カテゴリに対する重みで表現される。重み付け手法は、学習用データ

を用いて、単語の重みを計算する。具体的に用いられる手法は、テキスト内での重要語を求めるために最も一般的に用いられている tf・idf[6]やその応用[8]、学習用データから最小二乗法を用いて不正解が最小となるような重みを求める LLSF[12]などがある。これらの研究では、自らの提案する手法の優位性を述べるために他の手法との比較をおこなっているが、あくまでも手法の提案が主な目的である。カテゴリ表現以外では、単語の選択手法の提案[13]などがある。

2)一要素内、複数の手法の比較

Yang[14]は、5 つの単語の選択手法の比較を行っている。比較した手法は、文献の出現頻度、情報利得、相互情報量、χ^2 統計、term strength である。これらを 2 種類の学習用データを用いて、2 種類のカテゴリ表現において比較している。

3)複数の要素にわたる比較研究

Yang[15]は、カテゴリ表現、単語の選択を中心とした比較研究を行っている。カテゴリ表現手法は、最近傍法、線形最小二乗法、単語の単純照合手法の 3 つに関して新たに実験を行い、すでに論文で発表されている 11 種類の分類手法を含めた比較を行っている。さらに、最近傍法と線形最小二乗法を用いた分類では、情報利得、χ^2統計など 5 つの手法による特徴選択を行っており、分類精度が向上したことを示している。また、分類先の決定方法、評価方法などにも言及している。

Larson[7]は、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 通りの比較を行っている。石田[16),3)]は、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について比較を行っている。

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。

4　日本語テキスト特有の問題

日本語テキストは連続した文字列で表現されている。日本語テキストの自動分類を行う場合、英語テキストとは異なり、文章からの文字列の切り出しが、新たに必要な要素となる。文字列の切り出しは、切り出した文字列をもとにカテゴリ表現やテキスト表現をするので、分類に影響を与える大きな要素である。

文字列の切り出しには、文章を意味のある単語に切り出す形態素解析と文字列を数文字づつ切り出す N-gram などの方法がある。自然言語処理分野では、これらの手法の研究はさかんに行われ、充分に検討されている。

日本語テキストの自動分類において、文章から分類の手がかりとして何を用いているかということについて、既往研究を調べてみると、形態素解析を行った結果を用いている研究[17),18),20),21)]が多い。ただし、例えば、「形態素解析には juman を用いた」と単に記述されているだけであり、形態素解析を用いた理由に関する記述はなかった。

このように、自動分類の手がかりとなる文字列の切り出しは重要な要素であるにも関わらず、充分な検討がなされていないのが現状である。そのため、自動分類という観点から、意味を持つ単語がよいのか、単なる文字列でよいのかなど、分類の手かがりとして用いる場合の効果的な文字列の切り出し手法を検討する必要がある。

5　まとめ

既往研究で行われわれている各要素に関する研究をまとめると、表 2 のようになる。論文の中では、カテゴリ表現、テキスト表現、類似度計算などを含めて分類手法としているものがあり、テキスト表現手法、類似度計算

表2 既往研究で扱われている要素

	文献番号														
要素	2	3	7	8	10	11	13	14	15	16	17	18	19	20	21
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	△	△	△	△	△	○	○	○	○	△	△	△	○	△	○
カテゴリ表現手法	△	△	○	○	○	○	○	○	○	○	○	△	△	○	×
テキスト表現手法	△	△	○	△	△	○	×	×	×	×	△	△	△	△	×
類似度計算方法	△	△	△	△	△	×	×	×	×	△	△	△	△	△	×
分類先決定方法	△	△	△	○	○	×	×	△	○	△	△	○	△	△	×
評価尺度	△	△	△	△	△	△	△	△	○	△	△	△	△	△	△

＊○：手法の提案、比較研究が行われているもの，△：言及しているが、1手法のみ用いている，×：言及していない

方法などについて言及していないものもある。

表2から、全ての要素について、手法の提案、比較、また、提案や比較がされていなくても言及はされていることがわかる。特に、カテゴリ表現手法の提案や比較をしている研究が最も多い。これは、表1に示した自動分類に関わる要素が妥当であること、カテゴリ表現が中心的な問題であるということを示している。また、ある要素だけに注目した研究が多いことがわかる。

自動分類の流れを考えてみると、カテゴリ表現に、データの構造や量、文章からの文字列の切り出し、単語の選択は、非常に深く影響していることが考えられる。しかしながら、現状では、個々の要素に注目した研究が多い。そのため、今後の自動分類研究では、自動分類全体に関わる要素全体を視野に入れて、自動分類を捉えていかなければならない。また、手法の比較研究を行っている場合でも、単なる手法の比較で終わらせるのではなく、高い分類精度を示す手法はなぜそうなるのかなどを詳細に分析することが必要である。

【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) 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)

3) 石田栄美. "自動分類における分類メカニズム" 2000年度日本図書館情報学会春季研究集会発表要綱. pp.75-78(2000)

4) 岸田和明. 情報検索の理論と技術. 東京, 勁草書房, 1983.

5) 野本忠司. "確率モデルによる主題の自動抽出.", 情報処理学会研究報告, Vol.95, No.69 (NL-108) , pp.1-6(1995)

6) Salton, G., McGill, M.J. Introduction to Modern Information Retrieval. New York, McGraw-Hill, 1983.

7) 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)

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. " 情報処理学会自然言語処理 100-5, p.33-40(1994)

9) Maron, M.E. "Automatic indexing: An Experimental inquiry. " Journal of American computer Machinery. Vol.8, pp.404-417(1961)

10) 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)

11) 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)

13) Yang, Y., Wilbur W.J. "Using corpus statistics to remove redundant words in text categorization. " Journal of The American Society for Information Science. Vol.47, No.5, pp.357-369 (1996)

14) Yang, Y., Pedersen, J.P. "A Comparative Study on Feature Selection in Text Categorization. " Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97). pp.412-420(1997)

15) Yang, Y., "An Evaluation of statistical approaches to text categorization. " Journal of Information Retrieval. Vol.1, No. 1/2, pp.67-88(1999)

16) 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)

17) 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)

18) 西野文人. "テキスト分類のためのカテゴリ割り付け戦略. " 情報処理学会自然言語処理 106-3, p.13-18(1995)

19) 渡辺靖彦 他. "χ^2法を用いた重要漢字の自動抽出と文書の自動分類. " 情報処理学会情報学基礎 39-4, p.25-32(1995)

20) 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)

21) 持橋大地, 加来田裕和. "個人の選好に応じた単語の重要度の学習. " 情報処理学会自然言語処理 129-6, p.35-39(1999)