WWW探索支援のための記述意図によるテキスト分類

木村 託巳　山田 寛康　島津 明

北陸先端科学技術大学院大学 情報科学研究科

1 はじめに

インターネットの普及により，World Wide Web のページ数は日々増加を続けており，現在では，数十億にのぼるWebページが存在している．それに伴い，ユーザが必要な情報にたどり着くまでには，多くの時間を要するといった問題が起きている．このような問題から，テキスト分類は，重要なテキスト処理技術の一つとなった．ところがほとんどのテキスト分類研究は，主題に焦点をあてたものであり，テキストを一方向からしか見ていない．したがって，テキストを的確にとらえるために，異なった観点からもテキストを分類する必要がある．

本報告では,「意見」「説明」「紹介」といったコンテンツ作成者の意図という観点から，WWW上のテキストを分類することによって，効率的なブラウジングの支援を行うことを想定し，意図という観点から分類する手がかりには，どのような言語的情報が有効であるのか調べた結果を示す．

2 記述意図によるテキスト分類

Web閲覧者に「最近公開された映画の評判を聞きたい」という要求があったとする．その要求には「映画の批評や感想を述べたい」という意図を持った文書がマッチする．一方，一般的なテキスト分類では，同じ映画という分野の情報を見つけることは出来ても，それが宣伝を行っている文書なのか批評している文書なのか，という点は分からない．それは，ほとんどのテキスト分類が文章中に含まれる名詞や固有名詞となる単語，あるいはそれを抽象化したものを手がかりとして，分類を行っているためである．したがって，記述意図という観点からの分類には名詞以外の品詞となる単語の方が有効な手がかりになると推測することが出来る．本報告では，機械学習法を用いて，品詞が名詞となる単語の頻度を素性とした分類と品詞が名詞以外となる単語の頻度を素性とした分類の実験を行い，どちらが有効であるのか確かめた結果を示す．さらに，名詞以外のどの品詞に属する単語が有効であるのか，品詞ごとに素性から除いた実験について示す．

本報告では，記述意図のカテゴリとして以下のものを設定した．

- **意見:**
  ブックレビュー，感想，主張など，書き手が自分の考えを述べているもの．
- **説明・解説:**
  専門用語の説明，操作手順の説明など，読み手に深い理解をうながすもの．
- **紹介・案内:**
  人物紹介，製品紹介，交通経路案内など，読み手に知覚してもらうことを目的としたもの．
- **質問・回答:**
  疑問点・問題点の解決を求めているもの，およびその回答．(FAQ などもここに含める．)
- **報知・報告:**
  ニュース，レポートなど，読み手に何らかの事実を報せることを目的としたもの．
- **表現:**
  日記，随筆，小説など，書き手が何かを表現しているもの．
- **その他:**
  上記のどのカテゴリの条件にも当てはまらないもの．

これらのカテゴリには，意図というよりもむしろスタイルと呼ぶべきものも含まれているが，Web閲覧者の要求として，このようなスタイルの文書を目的とする場合もあるため，収集したデータ中に数多く見受けられるものは，カテゴリとして設定した．

3 実験条件

3.1 データセットの作成

様々な分野の文書を集めるために，Web上からランダムに300サイトのURLを取得し，そこを基点として16006ファイルのHTML形式で書かれた文書を収集．その中から，さらにランダムに1000ファイルを選び出し，そのうちの698ファイルに，先に述べたカテゴリにより，人手でラベル付けを行った．その内訳を表1に示す．ここで，残りの302ファイルは，日本語

表 1: カテゴリごとのファイル分布

カテゴリ	ファイル数
表現	85
質問・回答	46
意見	10
報知・報告	98
説明・解説	78
紹介・案内	300
その他	81
合計	698

表 2: 名詞に属する単語の頻度のみを素性とした実験

カテゴリ	precision	recall	F-measure
表現	65.7	63.2	64.4
質問・回答	77.7	73.7	75.6
意見	10.0	20.0	13.3
報知・報告	62.3	52.4	56.9
説明	61.7	46.4	53.0
紹介	73.0	65.8	69.2
その他	52.0	65.8	**58.1**
総合	57.5	55.3	56.4

表 3: 名詞以外に属する単語の頻度を素性とした実験

カテゴリ	precision	recall	F-measure
表現	69.0	72.2	**70.5**
質問・回答	87.1	79.1	**82.9**
意見	30.0	26.7	**28.3**
報知・報告	62.6	64.7	63.6
説明	63.7	47.2	**54.2**
紹介	72.7	76.5	74.5
その他	43.9	62.3	51.5
総合	61.3	72.2	**66.3**

表 4: 全ての単語の頻度を素性とした実験

カテゴリ	precision	recall	F-measure
表現	68.0	68.2	68.1
質問・回答	85.0	79.1	81.9
意見	30.0	26.7	**28.3**
報知・報告	64.0	64.9	**64.4**
説明	59.4	57.4	**58.4**
紹介	79.5	72.6	**75.9**
その他	50.4	64.6	56.6
総合	62.3	61.9	62.1

で書かれていない文書か，画像のみで全くテキストが含まれていないファイルであり，処理の対象外とした．

3.2 素性抽出

テキストを記述意図という観点から見たときの，言語的な特徴を明らかにすることに興味があるので，HTML タグの情報は利用しないこととし，HTML ファイルから全ての HTML タグを除去することで，プレーンなテキストに変換した．次に，茶筅 [1] を用いて形態素解析を行い，その結果から，文書中に現れる単語の頻度を計測し，機械学習法に用いる素性とした．

3.3 Support Vector Machine によるテキスト分類

ラベル付き訓練データから，記述意図による分類器を学習するために，SVM(Support Vector Machine)[2] と呼ばれる機械学習アルゴリズムを利用する．SVM には，汎化性能に優れており過学習を起こしにくく，多くの素性を扱うことが可能といった特徴がある．このような特徴は，非常に広範囲の分野にわたって文書が存在し，様々な調子で書かれているような WWW 上の文書を扱うのに都合が良い．ただし，SVM は本質的に二値分類器であるため，これを多値分類器とするために，one v.s. rest 法を用いる．また，SVM のカーネル関数に関して数回の予備実験を行った結果，linear カーネルが最も良い値を示していたので，以降の実験では，全て linear カーネルを利用する．そして 5-fold cross validation で実験を行う．

4 実験

4.1 実験 1: 名詞-対-名詞以外

記述意図という観点からのテキスト分類には，名詞以外の品詞の言語的情報の方が手がかりになると予想しているため，形態素解析の結果，名詞と判別された単語の頻度を素性とした実験と名詞以外の品詞と判別

表 5: 素性から除いたときの表 3 との差分（F 値）

	動詞	副詞	助動詞	助詞	感動詞	形容詞	記号	未知語	連体詞	接頭詞
表現	-3.3	+0.1	-6.1	-1.8	+0.1	-1.1	-1.1	-0.4	-0.3	-3.3
質問・回答	-1.0	0.0	-0.4	0.0	0.0	0.0	-4.5	-8.1	0.0	0.0
意見	-0.1	0.0	-8.3	+3.7	0.0	0.0	+3.7	-1.6	0.0	0.0
報知・報告	-5.5	-1.2	-6.3	-4.7	-2.0	-1.9	+1.6	-2.2	-3.3	-0.5
説明・解説	-6.5	-0.9	-7.0	-5.7	-0.7	-0.9	-0.8	-11.3	-0.7	-0.6
紹介・案内	-1.2	+0.2	-1.6	-2.7	+0.2	+0.1	-1.0	+0.8	+0.2	+0.1
その他	-0.6	-0.2	+1.3	+0.9	-0.5	+0.1	-5.4	-3.1	-0.3	-0.5
総合	-8.4	-5.4	-9.1	-6.4	-5.5	-5.7	-5.6	-8.4	-5.7	-5.2

された単語の頻度を素性とした実験を行った．(表 2, 表 3)

この結果は，SVM のパラメータを変化させ，F 値が最も高い値を示したときの結果である．以降の実験では，このときのパラメータを利用して実験を行っている．

表 2, 表 3 の結果によって，わずかな差ではあるが，確かに名詞以外の品詞となる単語を素性とした方が，より良い性能を示していることが分かる．そして，名詞に属する単語の頻度も素性に追加し，全ての単語の頻度を素性とした実験の結果 (表 4) より，総合的に見れば，名詞に属する単語の頻度を用いない方が，良い性能を示していることからも，当初の予測が確かめられた．

また，意見のカテゴリの結果が，低い値を示しているが，これは，事例が少なかったためだと考えられる．

4.2 実験 2：

実験 1 の結果を踏まえて，次に，名詞以外のどの品詞に属する単語が，記述意図の分類に貢献しているか調べるために，表 3 の実験環境から，一種類の品詞ごとに，素性として用いずに実験を行った．

表 5 は，それぞれ，抜いた素性と表 3 と比べたときの F 値の差分を示している．この表から，助動詞を素性から外したとき，総合結果の F-measure の値が-9.1 ポイントと最も低下し，継いで動詞が-8.4 ポイント，未知語が-8.4 ポイント，助詞が-6.4 ポイントの順に続く．即ち，記述意図の分類には，助動詞に属する単語が最も分類に貢献しており，動詞，未知語，助詞に属する単語も，分類に貢献していたことが分かる．

また，特筆すべきなのは，説明・解説カテゴリの分類で未知語を除いたときで，11.3 ポイントの低下が見られる．これは説明・解説のカテゴリに割り振ったデータの中に，専門用語の説明などが含まれていたため，未知語の影響が大きくなったと推測できる．

5 関連研究

Lee ら [3] は，獲得できる電子文書の増加と，テキストベースのアプリケーションの多様化から，テキストを主題以外の観点からも見ることが有効であると主張し，テキストのジャンル，スタイルといった観点からも分類を行っている．彼らの設定したカテゴリは，「Newspaper」「Review」「Research Paper」「Homepage」「Q&A」「Product Spec」であり，本研究で設定したカテゴリと近い構成になっているが全く同一とは言えないため，直接的な比較は出来ない．

6 まとめ

主題以外の観点として，記述意図という観点からテキストの分類にどのような素性が有効であるのか分析した結果を示した．記述意図という観点からの分類には，一般的なテキスト分類で用いられる手がかりとは異なる手がかりが有効であるという見方の可能性を示し，このような分類には，一般的なテキスト分類ではあまり用いられない助動詞，動詞，助詞といった品詞に属する単語の言語的な情報の方が貢献していることを明らかにした．

今後，分類に貢献した言語的な情報の中で，具体的にどの単語が有効であるのかを明らかにする．記述意図による分類が，効率的なブラウジングの支援に効果

的であるのか，心理実験を行って，確かめる．分類器の性能向上のために，訓練事例の少なさを補う方法か，効率よく訓練事例を増やす方法が必要である．

参考文献

[1] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸，形態素解析システム「茶筅」version 2.2.7 使用説明書，奈良先端科学技術大学院大学 松本研究室，2001.

[2] Vladimir N. Vapnik, Statistical Learning Theory, A Wiley-Interscience Publication, 1998.

[3] Young-Bae Lee & Sung Hyon Myaeng, Text Genre Classification with Genre-Revealing and Subject-Revealing F eatures, Proceedings of the 25th Annual International ACM SIGIR Congerence on R esearch and Development in Information Retrieval, pp145-150, 2002.