E-035

SVMを用いたブログ自動分類システム

Automatic Blog Classification System Using the Support Vector Machine

村井 基祐† 佐川 雄二† 田中 敏光†
Motohiro Murai Yuji Sagawa Toshimitsu Tanaka

1．はじめに

インターネットの普及に伴い Web 上では大量の情報が発信されている．近年では個人による情報の発信が特に盛んに行われるようになってきており，その代表的な情報の発信源の 1 つとしてブログ が挙げられる．ブログ は専用のツールやサービスを使用することで，簡単に記事を投稿・編集することが可能なことから，誰もが手軽に情報を発信することが出来る．ブログ記事には日記のほかに，ニュース記事に対する意見やコメントといったものや，商品に対する評価など，一般の人々の率直な意見が多く含まれている．これらの意見は関心のある話題に関する情報もしくはその手がかりとして，大変有用である．

その一方で，読み手の立場から必要な情報を含む記事を探すことを考えた場合，存在するブログ記事の量と記事の更新の早さから，検索対象は膨大な数になる．一般にブログ記事を検索する場合，ブログのポータルサイトが利用される．ポータルサイトでは収集されたブログ記事に対して，キーワードによる検索を行うことができるものや，サイトが用意したカテゴリに記事の投稿者がリンクすることによって，記事をカテゴリから絞り込むことができるものなどがある．

しかし，キーワードによる検索はユーザの技量によるところが大きいことや，投稿者によるカテゴリへのリンクは，すべてのブログサービスで行われている訳ではないことなど問題点は多い．

そこで本研究は，ブログ記事の検索を支援するため，ブログ記事を自動的に分類するシステムを開発することを目標とする．本稿では Support Vector Machine（SVM）を用いてブログ記事を分類する手法を提案し，実験の結果と考察について述べる．

2．提案手法

ブログ記事を分類する方法，自動分類のための学習・分類について述べる．

2.1. 分類方法

記事を探す際の手がかりとなる要素として，まずジャンルが挙げられる．記事の内容がどの分野に属するものであるかを特定することで，記事の探索対象を絞り込むことができる．しかし，ブログ記事の場合，たとえ同じジャンルであったとしても，投稿者の体験に基づいた日記のような記事もあれば，あるニュースに対するコメントを述べたような記事など，書き手の目的によってさまざまなスタイルが取られることがある．記事のスタイルによって得られる情報が異なるため，読み手であるユーザは，あるスタイルのみ，もしくはあるジャンルでも特定のスタイルの記事を必要とする場合も多い．

†名城大学 大学院 理工学研究科

そこで本研究では，ジャンルを特定すると同時に，記事がどのようなスタイルで書かれているかを特定することによって，記事を 2 つの観点から絞り込む方法を提案する．

2.1 ブログ記事の利用

自動分類には，予め特定するジャンルのカテゴリを用意しておき，ブログ記事がどのカテゴリに当てはまるかを判別する方法を用いる．分類には，カテゴリごとに収集された訓練データを用いてあらかじめ作成した分類器分類器を用いる．

分類器の作成に当たっては．分類済みの訓練データをどのように収集するかということが問題になる．ブログの自動分類の従来研究[1]では，分類があらかじめされているという点から，Web 上の掲示板に投稿された文書を訓練データとして利用し，掲示板でのカテゴリ構造を元に訓練データを与えている． そのほかの分類済みのデータとして，ニュース記事を利用することも考えられ．

しかし，本研究では直接ブログ記事を訓練データとして採用する．ブログで話題となる事柄は幅広いことや，生活に関わる内容など，ニュース記事だけでは網羅できないからである．また，ブログ記事は話題の移り変わりが早いので，ある期間で学習した分類器が，いずれ有効に作用しなくなるということが十分考えられる．よって，できるだけ新しい記事を訓練データとして使用することが好ましい．この点，ブログ記事は RSS などの機能により最新の記事を自動で入手しやすく，都合が良い．さらにポータルサイトによっては，投稿者がカテゴリを選択することによって記事に付与できるので，このような場合なら，分類済みの記事をそのまま訓練データとして利用することができる

2.2. 訓練データの収集

予めカテゴリを作成しておき，それに対応する正解データと不正解データを収集することによって訓練データを作成する．記事の収集方法はポータルサイトの RSS を監視し，更新された記事を取得することによって行う．これによって常に新しい記事を収集することができ，学習に最新のデータを反映させることができる．

多くのポータルサイトでは話題毎に RSS が作られており，これを利用してカテゴリ毎の記事を収集する．作成するカテゴリに当てはまる内容を扱っている RSS を正解データに指定し，逆に当てはまらない話題を扱っている RSS を不正解データに指定することによって，そのカテゴリが学習に必要な正解・不正解データを収集することが出来る．RSS 単位で必要な話題を選択することによって，柔軟なカテゴリの作成，新しい話題への対応といったことを可能としている．

2.3. 学習・分類

記事の本文を形態素解析器にかけて形態素解析を行い，得られた形態素から必要な品詞を取り出し，特徴語のリストを作成する．特徴語の出現頻度から重要度を計算し，ベクトルの重み付けに使用する．そして，訓練データごとに

特徴語のリストと重要度から文書ベクトルを作成し，これを SVM で学習することによって分類器を作成する．カテゴリ毎に選択された正解データ，不正解データを訓練データのセットとし，それぞれのカテゴリに当てはまるか否かを判別する分類器をカテゴリ毎に作成する．

分類では 1 つの記事に対して各カテゴリの分類器による判別を行い，正の結果が得られたカテゴリを記事に付与する．1 つの記事に対して複数のカテゴリが付与されることを認め，これによって 1 つの記事が複数の話題に及んでいるケースに対応させる．

3. 実験

ブログポータルサイトである goo ブログから収集した記事を用いて，ジャンルに対する分類実験を行った．

3.1. 実験内容

使用したカテゴリは 16 であり，goo ブログのカテゴリ構造を参考に作成している．各カテゴリ毎に正解データ，不正解データを収集し，それを訓練用データ，テスト用データに分けて実験を行った．表 1 に実験データ，表 2 に訓練データの諸元の一部を示す．

文書ベクトルの作成には，記事本文を形態素解析した結果から，助詞，助動詞を除いた内容語の語幹を使用した．語の重要度は TF-IDF を用いて算出し，重み付けを行った．テストデータも訓練データ同様に文書ベクトルを作成した．形態素解析器には MeCab0.95 を使用し，SVM による学習と分類には LibSVM2.84 を使用した．SVM はソフトマージンを用い，カーネル関数には線形カーネルを用いた．

3.2. 実験結果

それぞれのカテゴリの訓練用データで学習した分類器でテスト用データの判別を行い，精度（Accuracy），適合率（Precision），再現率（Recall）を求めた．表 3 に一部のカテゴリについての結果を示す．

表 1 ：実験データ

	全データ件数	正解データ件数	不正解データ件数
車・バイク	4,909	371	4,538
スポーツ	4,909	378	4,464
ゲーム	4,909	325	4,584
暮らし	4,909	322	4,587
お出かけ	4,909	269	4,640

表 2 ：訓練データ

	全データ件数	正解データ件数	不正解データ件数
車・バイク	7,085	3,022	4,083
スポーツ	7,616	3,600	4,016
ゲーム	6,748	2,623	4,125
暮らし	6,733	2,606	4,127
お出かけ	6,352	2,177	4,175

表 3 ：分類結果

	Accuracy	Precision	Recall
車・バイク	94.50	62.60	83.02
スポーツ	92.48	55.59	84.94
ゲーム	96.54	70.34	82.46
暮らし	88.88	33.28	69.25
お出かけ	92.04	37.13	65.43

3.3. 考察

全体的に精度は 84～97%，適合率は 28%～86%，再現率は 37～81%とカテゴリによって大きく異なった．「スポーツ」のような話題がはっきりしているカテゴリでは高い精度が得られ，「暮らし」のような内容にばらつきの大きいカテゴリでは高い精度は得られなかった．

適合率が低かったものの原因としては，今回使用した訓練データは 1 つの記事に対して 1 つのカテゴリしか付与されていなかったことが考えられる．ブログ記事の中には話題が複数に及んでいる記事も多いため，複数のカテゴリに対して正と判別した場合，不正解になってしまう．また，記事の書き手がカテゴリを選択する際に，相応しいカテゴリを選択していなかったため，本来正しく判別できているはずのものが不正解になってしまっている場合もある．

より精度の高い分類を行うためには，正解データ，不正解データに使用する RSS をカテゴリによって吟味して指定する必要がある．

また，スタイルの分類において，「日記」，「ニュース」，「評価・レビュー」に分類する簡易実験を行ったが，正答率は 50%程度しか得られなかった．訓練データ，テストデータのノイズや，特徴量の検討が十分で無かったことが原因と考えられる．

ノイズへの対策として，訓練データを手作業で収集することで，より正確な学習が可能になると思われる．また，品詞以外の情報をベクトルの特徴量として使用することによって，分類の精度を高める研究[2]も行われており，品詞以外の情報も積極的に取り入れることも検討している．

4. 今後の課題

今回行った実験では，分類器の性能を十分検証することが出来なかった．より多くのデータを収集することや，交差検定を行うなど，信頼度の高いテストを行う必要がある．また，今回のような機械的な判定ではなく，実際に人間が見た上での分類結果の評価も行う必要がある．

さらに今回作成したカテゴリは単位として大きく，より小さなカテゴリへも適用したいと考えている．

参考

[1]平野耕一，古林紀哉，高橋淳一，日本語圏ブログの自動分類，情報処理学会研究報告 Vol.2005 No.117(NL-170) pp.21-26，2005

[2] 川口敏広，松井藤五郎，大和田勇人，SVM と新聞記事を用いた Weblog からの意見文抽出，Vol.20th pp.1A3-3，2006