ブログ閲覧者の持つ多様な目的を考慮したブログ記事分類手法

鎌田　基之† 　横山　昌平†† 　福田　直樹†† 　石川　博††

† 静岡大学大学院情報学研究科　〒 432–8011 静岡県浜松市中区城北 3–5–1
†† 静岡大学情報学部情報科学科　〒 432–8011 静岡県浜松市中区城北 3–5–1
E-mail: †gs07016@s.inf.shizuoka.ac.jp, ††{yokoyama,fukuta,ishikawa}@inf.shizuoka.ac.jp

あらまし　ブログ記事の利用目的は，より詳しい情報源を得たい，著者の感想を得たいなど多様であることが予想される．これまで，収集したブログ記事集合により，ブログ記事から品詞やリンクなどの話題に依存しにくい統計情報を抽出し，C4.5 とナイーブベイズ，それら各々にバギングとブースティングを適用したもの，およびサポートベクターマシンでの，分類学習時の属性として用いることで，ブログ閲覧者の持つ多様な目的にそれぞれ特化した分類器の構築とそれらの評価を行ってきた．本研究では，相互関係性を考慮したブログ記事収集と，収集したブログ記事集合における目的別ブログ記事分類を実装し，ブログ閲覧者の持つ多様な目的にそれぞれ特化したブログ記事収集・分類システムを示す．

キーワード　ブログ，分類，サポートベクターマシン

A Goal Oriented Blog Entry Classification Method for Blog Readers

Motoyuki KAMADA†, Shohei YOKOYAMA††, Naoki FUKUTA††, and Hiroshi ISHIKAWA††

† Graduate School of Informatics, Shizuoka University　3–5–1 Jouhoku, Naka-ku, Hamamatsu-shi, Shizuoka, 432–8011 Japan
†† Department of Computer Science, Faculty of Informatics, Shizuoka University　3–5–1 Jouhoku, Naka-ku, Hamamatsu-shi, Shizuoka, 432–8011 Japan
E-mail: †gs07016@s.inf.shizuoka.ac.jp, ††{yokoyama,fukuta,ishikawa}@inf.shizuoka.ac.jp

Abstract　The actual motivations for searching, reading, and analyzing blogs are varied. One would do for locating rich information resources, and another one could do for investigating people's reactions for a specific event. Recently, we proposed an approach based on topic independent features of blog articles that can be statically gained from the number of anchor links or part of speeches. To realize topic independent classification of blog articles, we proposed machine learning based approach that uses C4.5, Naive Bayes and Support Vector Machine that can cover and adopt various needs of the readers. Also we showed the performance of these classifications. In this paper, we show our implementation for crawling strongly connected blog entries and classifying them for supporting various needs of blog readers.

Key words　Blog, Classification, Support Vector Machine

1. はじめに

オンライン上のニュースサイト（例えば，asahi.com）や企業の管理するポータルサイト（例えば，Yahoo! Japan）などに代表される既存のメディアとブログを比較した場合，ブログには幾つかの特徴があることが指摘されている [1]．例えば，ブログの内容には，オンライン上の日記という側面から特に個人的な意見や感想の表明をする要素が多く含まれ，更新の容易さにより時事性の高い内容に対して即時的な対応が見られやすい．ブログは，その著者のプロフィールとも深く結びついている可能性がある．ここでいうプロフィールとは，例えば，年齢や住んでいる土地や個人の関心事などである．

ブログ内で貼られるリンクによって作られる構造は，住んでいる土地の近さや現実での友人関係だけではなく，共通の関心事などの間接的な関係を反映した局限的なコミュニティを形成している可能性がある．ブログがもたらすコミュニティに関連して，中島らは，トラックバック利用状況の調査を行うことで，トラックバックリンクでつながったブログ記事の関係性を指摘している [2]．中島らは，トラックバックによる緩やかなコミュニティ形成を明らかにしており，ブログのコミュニティ発見に

対するトラックバックの重要性を指摘している．

我々は，文献 [3] で，トラックバックリンクを利用したクローリングをすることで，ある特定の話題に限定したブログ記事の集合を収集するための手法を提案した．さらに，話題の特徴語を用いたフィルタリングを行うことで，収集されるブログ記事集合が着目した話題に限定されることを示した．

我々は，文献 [4] で，文献 [3] の手法を用いて収集した，トラックバックリンクと特徴語を併用したクローリングによって得られたブログ記事の集合である**トラックバックコミュニティ**内において，ある話題についての情報を多くのリンクなどを用いて解説し，まとめている**解説型ブログ記事**，ある話題についての意見や感想を書いた**感想型ブログ記事**を定義し，その特徴を明らかにした．トラックバックコミュニティ内においてブログ記事が持つ，複数のブログ記事において共起したリンク数に基づいた分類手法の提案を行い，小規模な実験によりその効果を確かめた．

我々は，文献 [5] で，文献 [4] で定義した解説型ブログ記事についてブログ記事の書き手と読み手の持つ多様なニーズを考慮し，情報源となりうる文章を引用した**ニュース記事報告型**，情報源となりうる文章を引用し，著者自身の意見や感想も書かれている**ニュース記事意見追加型**の 2 つの小分類を定義した．トラックバックコミュニティ内においてブログ記事本文中から品詞数やリンク数などの統計情報を抽出し，学習アルゴリズムである，C4.5 とナイーブベイズを用いた分類学習の属性として用い，感想型ブログ記事を含めた 3 つの分類クラスへの分類を行った．同一話題での分類性能と学習済み分類器を学習対象とは異なる話題のブログ記事集合に用いた時の分類性能を評価し，抽出した属性の効果を確かめた．

ブログ記事の利用目的は，ユーザや目的により多様であることが予想される．ブログ記事を閲覧するという場面でも，例えば，ブログ記事の書き手と読み手というユーザの違いや，ジャーナリストとして閲覧するか，マーケティング目的で閲覧するかなどの目的の違いなどがある．ブログ上において議論が行われている時には，その意見の対立などの議論の構図を明らかにしたい場合，時系列に沿った意見の変遷を追いたい場合，ある特定の意見を持つ立場の人が何を情報源としているのかを分析したい場合がある．単一のブログ記事のみを見た場合にはこのような分析は困難であるが，ブログ記事間の関連に着目して，相互につながりを持ったブログ記事の集合として収集し，それらを，着目する観点ごとに分けることで，ブログ上で蓄積された様々な情報をより深く読み解くことが可能になるのではないかと期待される．

そのような，相互に強いつながりを持ったブログ記事集合から目的にあったブログ記事群を見つけ出す場合，それらの記事はほぼ同一時期に同一の話題について言及したものであることが多い．このような場合，語の出現頻度などを用いた単純な記事の分類手法を適用しても，明確に記事を分けることが難しい，あるいは，扱う話題に記事の分類手法が強く影響してしまうことから他の話題への適用が難しくなるなどの問題が生じることが考えられる．

本論文では，同一時期に同一の話題に対して言及されたブログ記事集合において，ブログ記事から品詞やリンクの統計情報を抽出し，それらを学習アルゴリズムの属性として用いることで，ブログ閲覧者の持つ,「より詳しい情報源を得たい」,「著者の感想を得たい」など多様な目的に特化した分類を実現する．各アルゴリズムの分類精度を評価し，他の分類手法との比較により本手法の有効性を示す．

本論文の構成は，次のようになっている．2. で，ある特定の話題に関連し同一時期に書かれたブログ記事集合の抽出手法について述べる．3. で，多様な目的に対応するフィルタの一例とそのための分類に用いる属性について述べ，学習アルゴリズムとして C4.5，ナイーブベイズ，それら各々にバギングとブースティングを適用したもの，およびサポートベクターマシンを用い，分類性能を評価する．4. で，それらを実装したシステムを示し，5. で，本研究と関連研究との差分を述べ，6. で，得られた成果をまとめる．

2. 同一時期に同一の話題に言及するブログ記事集合の抽出

本研究で対象とする同一時期に同一の話題に言及するブログ記事の収集は，トラックバックリンクを辿ることによって行う．シードブログ記事（収集の起点となるブログ記事）は，トラックバックを 1 つ以上受けているブログ記事とする．収集はトラックバックリンクを辿って行われ，辿るトラックバックリンクがなくなれば収集を終了する．ただし，シードブログ記事をもとに特徴語というものを設定し，その特徴語が本文に含まれるブログ記事のみを収集することとする．この方式を用いることによって，トラックバックスパムが収集されることを防ぐと共に，収集するブログ記事を，シードブログ記事で扱われている話題に絞り込むことが可能で，トラックバックスパムによるトラックバックコミュニティ内での話題の混濁が起こることを防ぐ．ここでのトラックバックスパムとは，起点のブログ記事の内容と無関係な内容が書かれたブログ記事のことである．また，通常トラックバックはブログ記事間で使用されるものであるため，あるブログ記事のトラックバック元であるトラックバックリンク先もまたブログ記事であると仮定できる．本クローリング手法ではクローリング中のウェブページがブログ記事であるかどうかを判定しない．なお，収集される情報はブログ記事の URL，HTML タグを含んだ本文，トラックバック数，トラックバック元のブログ記事 URL である．詳細は，文献 [4] を参照されたい．

3. ブログ記事の分類

3.1 フィルタの定義

検索エンジンやブログ検索エンジンを用いてキーワード検索を行った場合，キーワードが含まれ，そのキーワードに関連する話題を扱ったブログ記事集合を得ることが出来る．その場合，扱われている話題は同様でも，記事の内容や書かれ方など様々に異なる種類のブログ記事がそのブログ記事集合には含まれる．例えば，個人の日記として書かれたブログ記事，ブログ記事の

表 1　フィルタと分類クラス

フィルタ名	分類クラス
スパムフィルタ	スパムブログ記事 非スパムブログ記事
情報源フィルタ	情報源ブログ記事 非情報源ブログ記事
拡大情報源フィルタ	拡大情報源ブログ記事 非拡大情報源ブログ記事
感想フィルタ	感想ブログ記事 非感想ブログ記事

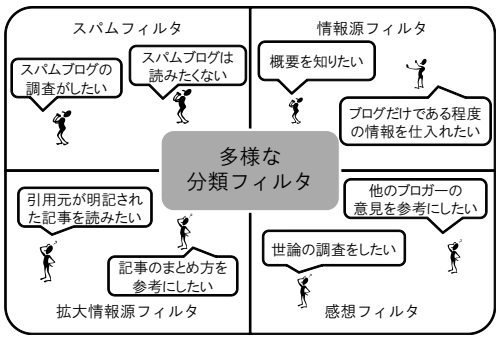

図 1　ブログ記事閲覧者の目的とそれに応じた分類方法の例

著者の考えが書かれた論説のようなブログ記事，ある話題について客観的事実を述べた二次情報源となりうるブログ記事，アフィリエイト収入を目的としたブログ記事，広告収入や他のサイトへの誘導のために自動生成されるスパムブログと呼ばれる本来のブログの目的からは逸脱したブログ記事などが挙げられる．同じキーワードで検索した場合，これらのブログ記事が混在して得られることとなり，キーワードには表しきれなかった本来の目的とは合致しないブログ記事を閲覧することもあり得る．我々は，ブログ記事の閲覧者のニーズを想定し，同一時期に同一の話題に言及するブログ記事を分類する 4 つのフィルタを定義した．そのフィルタと分類クラスを表 1 にまとめる．

これらのブログ記事の分類は，ブログ記事の閲覧者の目的に応じて推薦を行う場合に有効だと考えられる．その概要を図 1 に示す．ある話題についてすべてのブログ記事を閲覧したいが，スパムブログだけは避けたい場合はスパムフィルタによって，有益なブログ記事を得ることが出来る．逆に，スパムブログを調査したいと考える場合は，スパムフィルタによって必要なブログ記事のみを得ることが出来る．ブログ記事を情報の出所として閲覧している場合には，ニュース記事などの信頼できる記事の引用などを含むブログ記事を分類する情報源フィルタや，それらへのリンクを本文中に含むブログ記事を分類する拡大情報源フィルタによって，有益なブログ記事を得ることが出来る．ブログ記事からだけでは信頼がないと考える場合は，拡大情報源フィルタによって，より信頼できる情報源にたどり着くことも可能である．また，ブログ記事の著者にとっても，より信頼できる情報源やそれらを複数引用しまとめたブログ記事から整えられた文章が得られるため，自分の記事投稿時の参考とすることができる．ブログ記事を世論の一部と捉える場合，引用記事などは不必要であるため，感想フィルタを行うことで，ブログ記事の著者の感想や感情の記述を得ることが出来る．ブログ記事の著者たちの意見をまとめて記事を書きたい場合にも有益であろう．

3.1.1　スパムフィルタ

本研究では，同一時期に同一の話題に言及するブログ記事を収集するために，2. で述べたトラックバックと特徴語を併用した手法を用いている．そのため，内容に全く関連のないブログ記事やトラックバックスパムのほとんどを排除することができる．しかし，スパムブログには様々な種類が存在し，いくつかのスパムブログ記事が収集結果に含まれることがある．

本論文におけるスパムフィルタでは，スパムブログ記事と非スパムブログ記事の 2 クラスに分類する．本論文では，スパムブログ記事を次の（A）から（C）のいずれかに当てはまるものと定義する．(A）他のブログやニュース記事のスニペットなどの引用を羅列し記事にしているブログ記事．(B）記事の内容とは無関係のアフィリエイトリンクを大量に掲載しているブログ記事．(C）異なるブログサイトにおいて機械的に大量に投稿された同一内容のブログ記事．非スパムブログ記事は，上記のスパムブログ記事の定義に当てはまらないものと定義する．

3.1.2　情報源フィルタ

本論文での情報源フィルタでは，情報源ブログ記事と非情報源ブログ記事の 2 クラスに分類する．情報源ブログ記事は，ブログ記事本文中に情報源に相当する記述が存在するブログ記事と定義する．非情報源ブログ記事は，ブログ記事本文中に情報源に相当する記述が存在しないブログ記事と定義する．ここでの情報源とは，ある程度信頼できる発信元とし，例えば新聞社のポータルサイトや一般に公式サイトと呼ばれるサイトとする．情報源に相当する記述とは，情報源の文章の引用や，出典を明らかにした改変された情報源の文章などのこととする．

3.1.3　拡大情報源フィルタ

本論文での拡大情報源フィルタでは，拡大情報源ブログ記事と非拡大情報源ブログ記事の 2 クラスに分類する．拡大情報源ブログ記事は，ブログ記事本文中に情報源に相当する記述か，もしくは情報源に相当するサイトへのリンクが存在するブログ記事と定義する．非拡大情報源ブログ記事は，ブログ記事本文中に情報源に相当する記述と，情報源に相当するサイトへのリンクのいずれも存在しないブログ記事と定義する．情報源や情報源に相当する記述については，情報源フィルタにおける解釈と同様である．

3.1.4　感想フィルタ

本論文での感想フィルタでは，感想ブログ記事と非感想ブログ記事の 2 クラスに分類する．感想ブログ記事は，ブログ記事本文中にブログ記事の著者個人の感想や感情の記述，個人の日記にあたる記述が存在するブログ記事と定義する．非感想ブログ記事は，ブログ記事本文中にブログ記事の著者個人の感想や感情の記述と個人の日記にあたる記述のいずれも存在しないブログ記事と定義する．

3.2　分類に用いる属性

ブログ記事を分類するためには，複数の属性を抽出する必要がある．本研究では，ブログ記事の分野固有のキーワードなど

表 2　分類対象のトラックバックコミュニティが扱う話題と件数

話題	特徴語	収集件数
宮崎県知事選でそのまんま東氏当選	宮崎，知事 そのまんま，当選	104
自民党総裁選で福田新総理誕生	福田，総裁	104
ZARD の坂井泉水さん死去	坂井，泉水	195
光市母子殺害事件 元少年に死刑判決	光市，死刑	76
相棒-劇場版- 公開	相棒	69

表 3　データセットの内訳

フィルタ	クラス	話題				
		宮崎	総裁選	ZARD	死刑	相棒
スパム	スパム	1	1	13	9	3
	非スパム	103	103	182	67	66
情報源	情報源	46	56	60	39	17
	非情報源	57	47	122	28	49
拡大情報源	拡大情報源	66	72	99	53	32
	非拡大情報源	37	31	83	14	34
感想	感想	99	87	170	59	65
	非感想	4	16	12	8	1

に依存した分類を避ける目的で，記事内の特定の単語の出現頻度などを属性としては用いず，広い範囲のブログ記事に適用が可能となるような統計情報のみを用いることにした．本研究での分類に用いた属性は，(1) トラックバック数，(2) 相互トラックバック数，(3) 片方向トラックバック数，(4) 名詞，動詞，形容詞，副詞数，(5) 各品詞の TF・IDF 値の合計，(6) 各品詞の出現割合，(7) リンク数，(8) 共起したリンク数，(9) ニュース記事へのリンク数，(10) アフィリエイト数である．

各品詞の出現割合は，ブログ記事本文における名詞，動詞，形容詞，副詞の出現割合であり，式 1 を用いて求めた値である．pos には名詞 $noun$，動詞 $verb$，形容詞 adj，副詞 adv のいずれかが入り，ブログ記事毎に名詞，動詞，形容詞，副詞それぞれの出現数 C を計測し，それぞれを計測した 4 つの出現数の合計で割ったものである．

$$rate_{pos} = \frac{C_{pos}}{C_{noun} + C_{verb} + C_{adj} + C_{adv}} \tag{1}$$

その他の属性の詳細は，文献 [5] を参照されたい．

3.3　評価実験

3.3.1　学習アルゴリズムとデータ収集方法

本研究では，よく用いられ，効果をあげている学習アルゴリズムとして，C4.5 [6] とナイーブベイズ [7]（以後 NB と表記）とサポートベクターマシン [8]（以後 SVM と表記）を用い，各アルゴリズムごとの分類精度を評価した．また，C4.5 とナイーブベイズについては，複数の分類器の結果を統合するアンサンブル学習の手法であるバギング [9]（以後 Bagging と表記）とブースティング [10]（以後 Boosting と表記）を適用したものでも評価を行った．

本評価では，各アルゴリズムの実装として，ニュージーランドの Waikato 大学で開発が進められているデータマイニングツール Weka（Waikato Environment for Knowledge Analysis）[11] を用いた．学習アルゴリズムには，Weka に用意されている Quinlan の C4.5（C4.5）を実装した J48 とナイーブベイズ（NB）を実装した NaiveBayes とサポートベクターマシン（SVM）のライブラリである LibSVM [12], [13] を用いた．LibSVM のカーネルには，liner を用いた．バギング（Bagging）は Bagging，ブースティング（Boosting）は AdaBoostM1 を用いた．なお，それぞれの学習アルゴリズムにおける各種パラメータは Weka におけるデフォルト値を用いた．

今回，分類の対象となるブログ記事集合であるトラックバックコミュニティは，トラックバックを辿るクローラで収集を行った．収集したトラックバックコミュニティにおいて扱われている話題と収集の際に用いた特徴語と収集件数を表 2 に示す．なお，学習に用いるために，各データセットは，事前に手作業で分類を行っている．分類とそれぞれのクラスについては，3.1 で述べた通りである．なお，スパムフィルタ以外のフィルタの評価実験では，そのフィルタの性能を正確に測定するため，事前に手作業でスパムブログ記事と分類されたブログ記事を除いた集合に対してフィルタの適用を行う．データセットのフィルタ別の内訳を表 3 に示す．

本提案手法とは異なるアプローチに基づく分類手法として，湯浅らの手法 [14] がある．湯浅らの手法では，同一話題における単語間共起関係を用いて生成した特徴ベクトルを用い，文書を分類する．ただし，湯浅らの手法では，特徴ベクトルが話題に依存したものになる可能性がある．そこで，本提案手法の性質の 1 つである，分類器の話題への非依存性を確認するために，湯浅らの手法との比較を行う．ここでの特徴ベクトルの生成には，名詞のみを使用している．また，湯浅らの手法では学習の際に各クラスの典型的な記事を 3 つ選び学習を行うが，クラス内の記事数が 3 つに満たない場合は選択可能な最大数を学習対象とし，学習時に用いたブログ記事は分類性能評価時のテストデータには用いないようにした．

3.3.2　交差検定法による評価

表 2 のそれぞれのデータセットに対して，分類学習を行い，10 分割交差検定法で評価した．各実験条件における 10 分割交差検定法による分類精度の F 値を 3.1 で示した 4 つのフィルタ毎に平均をとり，まとめたものを表 4 に示す．比較対象として，データセットのクラス分布に従いランダムに分類した際の分類精度と湯浅らの手法を用いて学習・分類した際の分類精度も加えて示す．なお，表 4 中で，数値が太字となっている部分は，ランダムに分類した際の分類精度よりも劣ることを示す．

3.3.3　学習対象とは異なる話題への適用

学習済みフィルタを他の話題へ適用した場合の性能を評価するために，表 2 のそれぞれのデータセットに対して，3.3.2 で用いたものと同様の学習アルゴリズムで生成した分類器を用い，学習に用いたデータセット以外の 4 つのデータセットをそれぞれ分類し，その精度を計測する実験を行った．分類精度の F 値を 3.1 で示した 4 つのフィルタ毎に平均をとり，まとめたものを表 5 に示す．比較対象として，データセットのクラス分布に従いランダムに分類した際の分類精度も加えて示す．なお，表 5 中で，数値が太字となっている部分は，ランダムに分類した際の分類精度よりも劣ることを示す．

表 4　同一話題に対するフィルタの分類性能（F 値）

フィルタ	クラス	C4.5	NB	Bagging		Boosting		SVM	湯浅ら	ランダム
				C4.5	NB	C4.5	NB			
スパムフィルタ	スパムブログ記事	0.343	0.468	0.370	0.356	0.360	0.433	0.489	0.375	0.050
	非スパムブログ記事	0.978	0.979	0.982	0.977	0.978	0.983	0.981	**0.816**	0.950
情報源フィルタ	情報源ブログ記事	0.675	0.586	0.712	0.594	0.718	0.698	0.708	0.547	0.432
	非情報源ブログ記事	0.728	0.758	0.773	0.760	0.777	0.792	0.794	**0.550**	0.568
拡大情報源フィルタ	拡大情報源ブログ記事	0.821	0.693	0.839	0.718	0.836	0.744	0.835	**0.546**	0.632
	非拡大情報源ブログ記事	0.695	0.632	0.724	0.636	0.724	0.609	0.684	0.464	0.368
感想フィルタ	感想ブログ記事	0.953	0.955	0.959	0.956	0.958	0.953	0.942	**0.668**	0.921
	非感想ブログ記事	0.272	0.288	0.229	0.275	0.298	0.297	0.213	0.092	0.079

3.4　考　　察

3.4.1　学習アルゴリズムとフィルタ

表 4 のトラックバックを用いて収集した集合における同一話題に対するフィルタの分類性能を学習アルゴリズム別にみると，フィルタの種類によって特定のアルゴリズムによる分類性能が他より若干高くなる傾向が見られる．スパムフィルタのスパムブログ記事への分類精度は，NB と Boosting NB と SVM が他よりも 0.07 から 0.14 程度良い結果となっている．情報源フィルタの情報源ブログ記事と拡大情報源フィルタの拡大情報源ブログ記事への分類精度は，Boosting C4.5 と Bagging C4.5 と SVM が他よりも 0.1 程度良い結果となっている．ただし，それぞれの精度を F 値で見た場合で，スパムフィルタのスパムブログ記事への分類精度では 0.4 程度，感想フィルタの非感想ブログ記事への分類精度では 0.3 程度と，絶対的な性能としてはさほど高くないため，対象となるデータ集合ごとにアルゴリズムが期待通りの性能を出しているかどうかを確認しながら使い分ける必要があると考える．

3.4.2　学習対象とは異なる話題に対する適用可能性

表 4，表 5 より，学習対象とは異なる話題に対する適用性を評価した結果と同一話題で学習と分類を行い交差検定法で評価した結果とを比較すると，スパムフィルタや感想フィルタなど，わずかに精度を下げているフィルタが存在するものの，ほぼ同一話題で学習と分類を行い交差検定法で評価した結果と同じような精度で分類が可能であった．学習対象とは異なる話題に対しては，同一話題に対する分類とほぼ同程度の精度で分類が可能であり，本論文で提案する属性は適用可能である．また，話題に依存しない属性を用いている為，湯浅らの手法よりも良い精度で分類できていることを確認した．

4.　システムの実装

2. で述べた収集手法と 3. で述べた分類手法を実装したシステム “kuro” の概要を示す．Java と Javascript を用い Web アプリケーションとして実装した．学習アルゴリズムの実装としては，Weka API を用いた．

kuro は，ブログ記事の収集機構と分類機構から構成される．収集機構は，2. で述べたアルゴリズムを実装したものであり，シードブログ記事の URL と特徴語を入力することで，同一時期に同一の話題について書かれたブログ記事集合の収集を行う．

図 2　システムのブラウザ上での表示

シードブログ記事は複数指定可能である．収集したブログ記事は，保存し閲覧することができ，シードブログ記事からの距離（ID 順），トラックバック数順に並び替え可能である．

分類機構は，ブログ閲覧者の持つ目的に応じ，ブログ記事の分類を行う．ブログ閲覧者は，自身が持つ目的を明らかにするために，新たなフィルタと分類クラスを新規作成することができる．分類学習に必要なトレーニングデータの作成支援として，作成したフィルタ毎に分類クラスをブログ記事に付与することができるインタフェースを提供する．作成したトレーニングデータを用い，Weka API を利用し機械学習を行い分類器を生成する．学習アルゴリズムとしては，3.3.1 で示した全てのアルゴリズムが利用可能である．生成済みの分類器があれば，以後収集したブログ記事集合に対して分類器を適用し，作成した分類クラスに分類することができる．分類後は，指定した分類クラス毎にブログ記事を閲覧することも可能となる．実際のシステムのブラウザ上での表示を図 2 に示す．左フレームでシステムの操作を行い，右フレームで実際のブログ記事の閲覧を行う．

5.　関 連 研 究

ブログを対象とした研究には，ブログ記事の内容に焦点を当てたものと，ブログが作りあげる構造に焦点を当てたものがある．ブログ記事の内容に焦点を当てた研究として，大きく 2 種類の解析が存在する．ひとつが，ブログ記事の内容の時系列・空間解析である [1], [15], [16]．もうひとつが，ブログ記事，またはその著者の感情・感想のマイニングである [17]〜[20]．

これらの関連研究 [17]〜[20] において，機械学習で用いられている属性は，ブログ記事本文に含まれる単語である．属性となる単語の取捨選択については，情報利得やカイ二乗検定を用い

表 5　学習対象とは異なる話題に対するフィルタの分類性能（F 値）

フィルタ	クラス	C4.5	NB	Bagging		Boosting		SVM	湯浅ら	ランダム
				C4.5	NB	C4.5	NB			
スパムフィルタ	スパムブログ記事	0.134	0.124	0.122	0.101	0.135	0.076	0.268	0.055	0.050
	非スパムブログ記事	0.957	0.962	0.975	0.965	0.965	0.963	0.962	**0.467**	0.950
情報源フィルタ	情報源ブログ記事	0.642	0.586	0.655	0.587	0.647	0.614	0.689	**0.113**	0.432
	非情報源ブログ記事	0.653	0.747	0.716	0.753	0.730	0.737	0.737	0.573	0.568
拡大情報源フィルタ	拡大情報源ブログ記事	0.768	0.670	0.791	0.722	0.804	0.762	0.780	**0.229**	0.632
	非拡大情報源ブログ記事	0.641	0.634	0.609	0.640	0.650	0.616	0.659	**0.364**	0.368
感想フィルタ	感想ブログ記事	0.944	0.935	0.948	0.937	0.950	0.946	**0.920**	**0.607**	0.921
	非感想ブログ記事	0.136	0.188	0.130	0.169	0.159	0.175	0.230	**0.050**	0.079

るもの [19]，出現頻度の上位 N 単語までを用いるもの [17], [18]，ある閾値以上出現した単語のみを用いるもの [20] がある．

本研究では，ブログ記事の分野固有のキーワードなどに依存した分類を避ける目的で，上記の関連研究で用いられている記事内の特定の単語の出現頻度などを属性としては用いず，広い範囲のブログ記事に適用が可能となるような統計情報のみを用いることにした．そのため，本手法を分類時の前段や後段のフィルタとして用いることで，上記の関連研究と組み合わせて使用できる可能性がある．

6. おわりに

本論文では，同一時期に同一の話題について書かれたブログ記事集合において，話題に依存しにくい属性を用いた機械学習による分類手法を提案し，ブログ記事集合の収集と分類結果を閲覧するシステムの実装について述べた．

異なる分類目的において，他の手法やランダムに分類するよりも良い精度で分類することができた．統計情報に基づいた属性と機械学習を用いているため，本論文での分類とは異なる分類目的にもこの手法を適用可能な場面は多いと考えられる．一方で，本分類手法に基づく複数のフィルタ，あるいは本分類手法と他の関連する分類手法との組み合わせによる分類，およびその精度の解析については，今後の検討課題である．

今回示したシステムは，提案手法によるブログ記事の分類結果をブラウザ上で閲覧可能とするために実装したものである．ユーザビリティなどの評価については今後の課題である．

謝辞　本研究の一部は科学研究費補助金基盤研究（B）（課題番号 19300026），特定領域研究（課題番号 19024035）の助成による．

文　献

[1] Q. Mei, C. Liu, H. Su, and C. Zhai, “A probabilistic approach to spatiotemporal theme pattern mining on weblogs,” WWW ’06: Proceedings of the 15th international conference on World Wide Web, pp.533–542, 2006.

[2] 中島伸介，舘村純一，原良憲，田中克己，植村俊亮，“ブログ空間におけるトラックバック利用状況の調査および考察,” 日本データベース学会論文誌（DBSJ Letters），vol.5，no.1，pp.17–20，june 2006.

[3] 鎌田基之，福田直樹，石川博，“Trackback と特徴語に基づく blog クローリングと blog 記事の推薦,” 電子情報通信学会第 18 回データ工学ワークショップ論文集，A8-4，2007.

[4] 鎌田基之，戸田智子，黒田晋矢，福田直樹，石川博，“トラックバックコミュニティにおける特徴的なブログ記事集合の抽出について,” 電子情報通信学会技術研究報告 DE2007-64，vol.107，no.131，pp.253-258，2007.

[5] 鎌田基之，福田直樹，石川博，“トラックバックコミュニティにおける特徴的なブログ記事集合の抽出について,” 電子情報通信学会第 19 回データ工学ワークショップ論文集，B4-3，4 2008.

[6] J.R. Quinlan, C4.5: programs for machine learning, Morgan Kaufmann Publishers Inc., 1993.

[7] P. Langley, W. Iba, and K. Thompson, “An analysis of bayesian classifiers,” In Proceedings of the Tenth National Conference on Artificial Intelligence, vol.228, pp.223–228, 1992.

[8] C. Cortes, and V. Vapnik, “Support vector networks,” Machine Learning, pp.273–297, 1995.

[9] L. Breiman, and L. Breiman, “Bagging predictors,” Machine Learning, pp.123–140, 1996.

[10] Y. Freund, and R.E. Schapire, “Experiments with a new boosting algorithm,” In Proceedings of the Thirteenth International Conference on Machine Learning, pp.148–156, Morgan Kaufmann, 1996.

[11] I.H. Witten, and E. Frank, Data Mining: Practical machine learning tools and techniques, Morgan Kaufmann Publishers Inc., 2005.

[12] R.E. Fan, P.H. Chen, and C.J. Lin, “Working set selection using second order information for training support vector machines,” J. Mach. Learn. Res., vol.6, pp.1889–1918, 2005.

[13] C.C. Chang, and C.J. Lin, LIBSVM: a library for support vector machines, , 2001, Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[14] 湯浅夏樹，上田徹，外川文雄，“大量文書データ中の単語間共起を利用した文書分類,” 情報処理学会論文誌，no.8，pp1819-1827，vol.36，no.8，pp.1819–1927，August 1995.

[15] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins, “Information diffusion through blogspace,” WWW ’04: Proceedings of the 13th international conference on World Wide Web, pp.491–501, 2004.

[16] N. Glance, M. Hurst, and T. Tomokiyo, “Blogpulse: Automated trend discovery for weblogs,” WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, vol.2004, 2004.

[17] G. Leshed, and J.J. Kaye, “Understanding how bloggers feel: recognizing affect in blog posts,” CHI ’06: CHI ’06 extended abstracts on Human factors in computing systems, pp.1019–1024, 2006.

[18] G. Mishne, “Experiments with mood classification in blog posts,” Style2005–1st Workshop on Stylistic Analysis of Text for Information Access, at SIGIR 2005, 2005.

[19] X. Ni, G.R. Xue, X. Ling, Y. Yu, and Q. Yang, “Exploring in the weblog space by detecting informative and affective articles,” WWW ’07: Proceedings of the 16th international conference on World Wide Web, pp.281–290, 2007.

[20] 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 2005-NL-170，2005.