情報処理学会論文誌　データベース　Vol. 2　No. 2　56–70 (June 2009)

ブログの相互関係性を考慮したブログ記事分類手法の検討

鎌田基之[†1,*1]　福田直樹[†2]
横山昌平[†2]　石川博[†2]

ブログの増加にともない，ブログからの効果的な情報の収集は重要な課題となっている．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．ブログ記事は，それ自体を単体の独立したウェブページとしてとらえるのではなく，トラックバックなどによる相互関係性を持った集合としてとらえた場合に，ある話題に対する見解の相違の様子などのような，単体のブログ記事からは得にくい有益な情報が得られる場合が考えられる．本論文では，ブログ特有のトラックバックなどによるブログ記事相互の関係性を考慮して収集したブログ記事集合に対し，そこで形成されたコミュニティの特性および品詞やリンクなどの記事内の話題に影響されにくい統計情報を分類学習時の属性として用いる手法を提案する．本手法によって得られた分類器を，学習データとは異なる話題に対する分類問題に適用したときの性能を評価し，学習アルゴリズムや分類目的によって多少異なるものの，ブログの持つコミュニティとしての性質を保って収集されたブログ記事集合に対する分類精度が高く，学習に用いた話題に依存しにくいという特性を持つことを示す．

A Classification Method for Strongly Connected Blog Entries

Motoyuki Kamada,[†1,*1] Naoki Fukuta,[†2]
Shohei Yokoyama[†2] and Hiroshi Ishikawa[†2]

Blogosphere is evolving as an important information resource and therefore there is a certain need to realize blog-specific effective search engines, crawlers, and article classifiers. We can find blog entries for a certain event by using blog search engines with ordinary ranking algorithms. It is difficult to find out a set of blog entries that have meaningful relations among them. The actual motivations for looking 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. In our approach, we use topic independent features of blog articles that can be statically gained from the number of anchor links or part of speeches to realize a classification using various machine learning algorithms that can cover and adopt various needs of the readers. We show that the performance of classification does not deeply depend on the topic of articles in the learning set.

1. まえがき

平成 20 年 7 月の発表では，平成 20 年 1 月現在約 1690 万のブログが日本国内で開設され，記事数は約 13 億 5000 万件にのぼるとしている．1 カ月に 1 回以上記事が更新されているアクティブブログは，約 300 万となっている．ブログの増加にともない，この重要な情報源を対象とした研究や商用サービスが提案されつつある．ブログに特化したサーチエンジンの開発をはじめとして，ブログからの効果的な情報の収集は重要な課題となっている．

オンライン上のニュースサイト（たとえば，asahi.com）や企業の管理するポータルサイト（たとえば，Yahoo! Japan）などに代表される既存のメディアと比較した場合，ブログにはいくつかの特徴があることが指摘されている[1)]．たとえば，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付いている可能性がある．ここでいうプロフィールとは，たとえば，年齢や住んでいる土地や個人の関心事などである．

ブログ内で張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．

†1 静岡大学大学院情報学研究科
Graduate School of Informatics, Shizuoka University
†2 静岡大学情報学部情報科学科
Department of Computer Science, Faculty of Informatics, Shizuoka University
*1 現在，京セラ株式会社
Presently with KYOCERA Corporation

本論文では，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に対して，それらがトラックバックなどで形成する相互の関係性を維持しながら収集できたときに，それらを効果的に分類するための手法を実現することを考える．

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，扱う話題の特性が記事の分類学習結果に強く影響してしまうことから過学習状態となり，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予想される．

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を対象に，ブログ特有のトラックバックなどによるブログ記事相互の関係性を考慮して収集したブログ記事集合に対し，そこで形成されたコミュニティの特性および品詞やリンクなどの記事内の話題に影響されにくい統計情報を分類学習時の属性として用いることにより，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」など多様な目的にそれぞれ特化した分類器を，分類対象となるブログ記事の扱う話題に強く依存しない形で実現するための手法を提案する．本手法による分類器の性能を，単語の出現頻度に基づく分類学習手法との比較，および対象とするブログ記事集合内のコミュニティ形成の度合いの違いの 2 つの側面から比較し，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 章で，本論文でのブログ記事分類の目的と前提条件を示した後，多様な目的に対応するブログ記事分類（フィルタ）の一例とブログ記事集合が形成するコミュニティを生かしたブログ記事分類手法について述べる．3 章で，本提案手法の評価方法およびその実験条件を示し，本提案手法の分類性能の特性を評価する．4 章で，本研究と関連研究との差分を述べ，5 章で，得られた成果をまとめる．

2. ブログ記事の分類

2.1 ブログ記事分類の目的と前提条件

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，たとえば，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ときには，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．ブログ記事に対する分析手法には，クラスタリングなどの技術を用いてブログ全体の話題の傾向を分析する手法も提案されている[3)]．本論文では，むしろ収集したブログ記事の話題の傾向は予測可能か既知のものである場合での，ブログ記事の効果的な分類手法を考える．

ブログの収集は，すでに多くの商用検索エンジンによって行われつつあり，その収集結果を利用したブログ専用の検索エンジンも実現されつつある[4)]．ブログ検索エンジンを用いれば，そのランキングアルゴリズムを通して，ある特定の話題に関連したブログ記事を単体で発見することは可能である．しかしながら，ブログ検索エンジンでのランキング上位 N 件を取得する方法では，ある種のつながりを持ったブログ記事を 1 つの集合として収集するには，適さない場合がある．本論文では，何らかの手段により，トラックバックなどで形成する相互の関係性を維持しながらブログ記事集合を収集することが可能であると仮定する．本論文では，そのように収集されたそれぞれのブログ記事集合のことを，トラックバックコミュニティと呼ぶこととする．

本論文では，ある特定に話題に関連して同一時期に書かれたブログ記事集合を，トラックバックなどで形成される相互の関係性を維持しながらトラックバックコミュニティとして収集できることを仮定する．そのように収集できたブログ記事集合に対して，それらがコミュニティを形成するための手段として用いられるトラックバックなどの性質を効果的に活用することにより，それらのトラックバックコミュニティ内から，ブログ記事分析者が持つ多様な目的に合致するようなブログ記事を効果的に分類する方法について考える．

2.2 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できる．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．たとえば，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

表 1　フィルタと分類クラス
Table 1　Filters and their classifications.

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

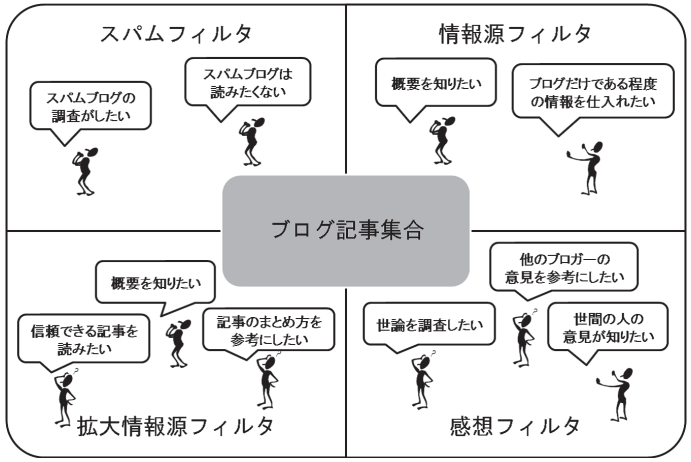

図 1　ブログ記事閲覧者の目的とそれに応じた分類方法の例
Fig. 1　Reader-centered blog article classification.

のブログの目的からは逸脱したブログ記事などがあげられる．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うる．

我々は，ブログ記事の閲覧者のニーズを想定したブログ記事分類方法の例として，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[5]．そのフィルタと分類クラスを**表 1** にまとめ，その概要を**図 1** に示す．

表 1 の 4 つのフィルタでは，たとえば，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りスパムブログをフィルタリングすることで，有益なブログ記事を探し出すことが容易になる．逆に，スパムブログそのものを調査対象と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できる．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できる．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，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ためのブログ記事を得ることができる．ブログ記事を世論の一部ととらえる場合，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できる．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利用可能と考えられる．本論文では，表 1 に示す 4 つのフィルタをブログ記事分類問題の例題として，以降のブログ記事分類手法の有効性の評価に用いる．

2.2.1　スパムフィルタ

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．ここで，本論文では，スパムブログ記事を次の（A）から（C）のいずれかにあ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自動的に取得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あてはまらないものと定義する．

スパムブログ記事は，一般的にノイズとして扱われ，それらを判別する研究もなされている[6]．スパムブログ記事を分類することは，一般的なブログ記事の読み手には有効なフィルタリング機能の 1 つとして利用される．また，分類結果は，スパムブログ記事を対象に研究や調査を行いたい場合にも利用可能である．

2.2.2　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，

表 2　分類に用いる属性
Table 2　Features used in the classification.

属性	説明
トラックバック数	ブログ記事が受けているトラックバックの合計
相互トラックバック数	ブログ記事がトラックバック元と相互にトラックバックしあっているトラックバックの数
片方向トラックバック数	ブログ記事がトラックバック元から一方的に受けているトラックバックの数
名詞，動詞，形容詞，副詞数	ブログ記事本文の名詞，動詞，形容詞，副詞数
各品詞の TF・IDF 値の合計	ブログ記事本文の名詞，動詞，形容詞，副詞の TF・IDF 値をそれぞれ合計したもの
各品詞の出現割合	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
リンク数	ブログ記事本文に含まれるリンクの数
共起したリンク数	ブログ記事集合中の 2 記事以上で共起したリンクの所持数
新聞社・通信社の記事へのリンク数	ブログ記事本文中における新聞社・通信社へのリンク数（15 サイト）
アフィリエイト数	ブログ記事本文中におけるアフィリエイトリンク数（8 サイト）

たと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．

ブログ記事のみから情報を得ようとする場合，ある程度信頼できるブログ記事を見つけたいというニーズが考えられる．その場合，著者個人の感想が書かれたブログ記事は必要ではない．情報源フィルタにより，ある話題について詳細な情報が記述されたブログ記事だけを閲覧することができる．

2.2.3　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．

ブログ記事はそれほど信頼していないが，信頼できる情報源にたどり着くための手段として用いたい場合，ブログ記事の読み手にとっては，ブログ記事本文に含まれるリンクは重要な要素である．またブログ記事の書き手にとっても，同じ話題を題材にブログ記事を書きたい場合に，それらのブログ記事を参考にすることで，より信頼できる情報源の情報を引用し，リンクすることができる．

2.2.4　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．

ブログ記事からの評判情報の抽出に関する研究やサービスは多数存在する．ブログ記事には情報源の文章を引用したものもあり，評判情報や世間の反応などを知りたい場合には必要ない．著者個人の感想や感情の記述が書かれているブログ記事を分類することで，世論の調査や評判情報の抽出の補助ができると考える．

2.3　話題依存性の低い分類器学習手法

本論文では，機械学習における分類学習手法を用いてブログ記事を分類することを考える．ブログ記事を分類学習手法によって適切に分類するためには，ブログ記事から分類に必要な複数の属性値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分類学習のための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を**表 2** にまとめる．

トラックバック数は，ブログ記事が受けているトラックバックの合計数である．相互トラックバック数は，相互トラックバック数を計測したものである．トラックバックを受けたトラックバック先のブログ記事から，トラックバック元のブログ記事へトラックバックする

ことによって，お互いの記事をトラックバックにより自動的に作られるリンクで相互にリンクし合うということが行われており，これを相互トラックバックと呼ぶ．片方向トラックバック数は，相互トラックバックの状態になっていないトラックバックの数を計測したものである．トラックバック数から相互トラックバック数を除いて求める．

名詞，動詞，形容詞，副詞数は，ブログ記事の本文中に含まれる名詞，動詞，形容詞，副詞の種類数である．複数回以上出現したとしても 1 として計測している．

各品詞の TF・IDF 値の合計は，ブログ記事本文に出現する各品詞の TF・IDF 値をもとに式 (1) を用いて求めた値である．式 (1) 中の，pos には名詞，動詞，形容詞，副詞のいずれかが入り，$tfidf_k$ はブログ記事本文中に出現したある品詞 pos_k の TF・IDF 値，k はブログ記事本文中の各品詞数である．

$$tfidfsize_{pos,m} = \sqrt{\sum_{n=1}^{k} {tfidf_n}^2} \tag{1}$$

トラックバックコミュニティ内では，同一の話題に言及したブログ記事のみが存在する．その集合の中で，各ブログ記事が持つ各品詞について TF・IDF 値を計算し，ブログ記事ごとに合計を求めることで，ブログ記事の独自性を測ることができると考える．あるブログ記事で出現頻度が高く，その他のブログ記事では出現頻度が低い単語が多く含まれるほど $tfidfsize$ は大きくなりやすい．そのため，同じ話題を扱っているブログ記事集合において，このブログ記事がどれだけ独自の単語を用いているかを計ることができると考えられる．すなわち，その値が大きい方がブログ記事の著者自身の言葉，すなわち感想や意見にあたる文章が多く含まれ，小さい方が情報源に相当する記述が多く含まれるのではないかと予測できる．

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 (2) を用いて求めた値である．pos には名詞 $noun$，動詞 $verb$，形容詞 $adjective$，副詞 $adverb$ のいずれかが入り，ブログ記事ごとに名詞，動詞，形容詞，副詞それぞれの出現数 $Count$ 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．

$$rate_{pos} = \frac{Count_{pos}}{Count_{noun} + Count_{verb} + Count_{adjective} + Count_{adverb}} \tag{2}$$

リンク数は，ブログ記事本文中に含まれるリンクの数である．HTML のアンカータグ中の href 属性において「http」から始まるものをリンクと判断し，その数を計測している．

共起したリンク数は，トラックバックコミュニティ内で 2 記事以上で共起したリンクの数である．共起したリンクについては，文献 7) で，トラックバックコミュニティ内で共起するリンクは，話題の情報源となりうるウェブページへのリンクであることが多いため，分類の指標としての効果が確かめられている．また，アフィリエイトなどのリンクが大量に存在することで，情報源へのリンクがまったく存在しなくても共起するリンクを持つことが起こる可能性がある．このことがノイズとなり分類結果に影響を与えると考えられるため，ブログ記事が持つ共起するリンクの抽出においては，著名なアフィリエイト，ブログランキング，ブログサービス固有のタグに含まれる URL をチェックすることで，それらを無視することとしている．

新聞社・通信社の記事へのリンク数は，ブログ記事本文中における新聞社・通信社へのリンク数である．新聞社・通信社とは，Yahoo!や MSN などのポータルサイトや読売新聞や朝日新聞などが運営するニュースサイトのことである．これらのサイトは，特に情報源として信頼できると考えられることから抽出することとした．抽出する新聞社・通信社の URL は計 15 サイトである．

アフィリエイト数は，ブログ記事本文中におけるアフィリエイトリンク数である．アフィリエイトリンクは話題と関連する商品へのリンクであることが考えられ，特徴的な要素となりうるため抽出することとした．抽出するアフィリエイトサイトは計 8 サイトである．

3. 評 価 実 験

3.1 分類対象となるブログ記事の収集方法と学習アルゴリズム

本論文では，2 種類の方法で収集したブログ記事集合を対象に，本手法における分類手法を適用し，その性能を評価する．ブログ記事の収集方法の 1 つはトラックバックを用いた方法であり，もう 1 つは Google ブログ検索を用いた方法である．

分類の対象となるブログ記事集合にトラックバックコミュニティとしての性質を持たせるために，ブログ記事集合の収集をトラックバックに基づくフォーカストクローラ[7] を用いて収集を行った．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たどって行われ，たど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ともに，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

```
アルゴリズム function crawling( Blog b, Topic t )
if( b の本文に t が含まれる )
    b の HTML を解析し，すべての TrackBack 元を抽出する;
    解析した情報を格納する;
    foreach tb_origin(TrackBack 元集合)
      if( tb_origin が未解析 )
        crawling(tb_origin, t);
      end if
    end foreach
else
    b を解析済みとする;
end if
```

図 2　クローリングアルゴリズム
Fig. 2　Crawling algorithm.

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．本アルゴリズムを**図 2** に示す．本アルゴリズムは，単純な条件付き深さ優先クローリングである．しかしながら，収集の過程で比較的効率良くスパム記事を排除でき，収集対象の絞り込みや収集にかかる時間をコントロールすることが容易であるという特長を持つ．

収集したトラックバックコミュニティが扱う話題は，「宮崎県知事選でそのまんま東氏当選」「自民党総裁選で福田新総理誕生」「ZARD の坂井泉水さん死去」「光市母子殺害事件 元少年に死刑判決」「相棒-劇場版-公開」の 5 つである．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 3** 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，2.2 節で述べたとおりである．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 4** に示す．

本論文で提案する分類手法そのものは，その対象はトラックバックを用いて収集したブログ集合に限定されるわけでなく，同一時期に同一の話題に対して言及されたブログ記事であれば，理論上はどのような集合に対しても適用可能である．データの収集方法に起因するブログ集合内でのトラックバックコミュニティの形成度の違いが分類精度に与える影響を調べるために，Google ブログ検索を用い，トラックバックを用いて収集したブログ記事集合と同様の 5 つの話題のブログ記事集合を収集したものも，別途評価対象として用意した．収集の際に用いた検索条件および収集件数を**表 5** に示す．Google ブログ検索からは最大で 999 件のブログ記事へのリンクが取得可能であるが，実験対象となるブログ記事本体を得るためには，ブログ記事そのものにアクセスし HTML タグを含む本文を取得する必要がある．本実験で Google ブログ検索から実際に収集できたブログ記事の件数は，表 5 に示す件数となった．なお，トラックバックで収集したものと同様に，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，2.2 節で述べたとおりである．データセットのフィルタ別の内訳を表 4 に示す．属性については，Google ブログ検索により収集されたブログ記事は，その出所が多岐にわたっており，自動的なトラックバックの抽出が困難であったため，表 2 に示す属性から，トラックバックに関する属性である，トラックバック数，相互トラックバック数，片方向トラックバック数を除いた属性を用いた．

表 3　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
Table 3　Topics and the number of articles for each trackback community.

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

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[8] とナイーブベイズ[9] とサポートベクターマシン[10] 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[11] とブースティング[12] 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）[13] 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

表 4　データセットの内訳
Table 4　Details of the dataset.

収集方法	フィルタ	クラス	話題 宮崎	総裁選	ZARD	死刑判決	相棒
トラックバック	スパムフィルタ	スパムブログ記事	1	1	13	9	3
		非スパムブログ記事	103	103	182	67	66
	情報源フィルタ	情報源ブログ記事	46	56	60	39	17
		非情報源ブログ記事	57	47	122	28	49
	拡大情報源フィルタ	拡大情報源ブログ記事	66	72	99	53	32
		非拡大情報源ブログ記事	37	31	83	14	34
	感想フィルタ	感想ブログ記事	99	87	170	59	65
		非感想ブログ記事	4	16	12	8	1
Google ブログ検索	スパムフィルタ	スパムブログ記事	13	42	5	7	27
		非スパムブログ記事	317	183	457	259	389
	情報源フィルタ	情報源ブログ記事	93	85	123	167	63
		非情報源ブログ記事	224	98	334	92	326
	拡大情報源フィルタ	拡大情報源ブログ記事	142	94	221	177	155
		非拡大情報源ブログ記事	175	89	236	82	234
	感想フィルタ	感想ブログ記事	289	149	436	226	379
		非感想ブログ記事	28	34	21	33	10

表 5　Google ブログ検索から収集したブログ記事集合
Table 5　Blog articles collected from Google Blog Search.

話題	キーワード	期間	収集件数
宮崎県知事選でそのまんま東氏当選	そのまんま，宮崎 知事，当選	2007-01-21 から 2007-01-28	330
自民党総裁選で福田新総裁誕生	福田，総裁選	2007-09-23 から 2007-09-30	225
ZARD の坂井泉水さん死去	坂井，泉水	2007-05-26 から 2007-05-31	462
光市母子殺害事件 元少年に死刑判決	死刑，光市	2008-04-22 から 2008-04-29	266
相棒-劇場版-公開	相棒，劇場版	2008-04-29 から 2008-05-13	416

ターマシン（SVM）のライブラリである LibSVM[14),15)] を用いた．LibSVM のカーネルには，linear 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．

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，Ni らの手法[16)] がある．Ni らの手法では，ブログ記事を対象とし，ブログ記事に出現する単語そのものをサポートベクターマシンの属性として用いる．ただし，Ni らの手法では，単語そのものを属性として用いているため，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，Ni らの手法との比較を行う．Ni らは，属性選択には情報利得を用い，全単語の 70% を用いると最も分類精度が良いとしている．ただし，本評価実験で用いたデータセットにおいて，情報利得 IG を計算したところ，それぞれのフィルタにおいて，90% 程度の属性が $IG \leq 0$ となった．そのため，本評価実験では，情報利得を用いて属性選択を行ったもの，属性選択を行わずすべての属性を用いたものの 2 種類について実験を行うことにする．また，Ni らは，サポートベクターマシンのカーネルとして RBF カーネルを用いているが，本論文での提案手法では，3.2 節以降で述べるとおり，linear カーネルを用いた場合に比較的良い精度を上げていることから，linear カーネルと RBF カーネルの 2 つのカーネルそれぞれを用いた場合について実験を行う．すなわち，属性選択を行わずサポートベクターマシンのカーネルとして linear カーネルを用いたもの（linear），属性選択を行わずサポートベクターマシンのカーネルとして RBF カーネルを用いたもの（RBF），情報利得を用いて属性選択を行いサポートベクターマシンのカーネルとして linear カーネルを用いたもの（IG&linear），情報利得を用いて属性選択を行いサポートベクターマシンのカーネルとして RBF カーネルを用いた

表 6　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（F 値）
Table 6　Performance of classification filters applied to the learned events (F-measure).

フィルタ	クラス	C4.5	NB	Bagging C4.5	Bagging NB	Boosting C4.5	Boosting NB	SVM	ランダム
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	0.050
	非スパムブログ記事	0.978	0.979	0.982	0.977	0.978	0.983	0.981	0.950
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	0.432
	非情報源ブログ記事	0.728	0.758	0.773	0.760	0.777	0.792	0.794	0.568
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	0.632
	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	0.368
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	0.921
	非感想ブログ記事	0.272	0.288	0.229	0.275	0.298	0.297	0.213	0.079
平均		0.683	0.670	0.698	0.659	0.706	0.689	0.706	0.500

表 7　Google ブログ検索を用いて収集した集合における同一話題でのフィルタの分類性能（F 値）
Table 7　Performance of classification filters on Google-based crawled articles applied to the learned events (F-measure).

フィルタ	クラス	C4.5	NB	Bagging C4.5	Bagging NB	Boosting C4.5	Boosting NB	SVM	ランダム
スパムフィルタ	スパムブログ記事	0.177	0.176	0.205	0.184	0.217	0.172	0.104	0.066
	非スパムブログ記事	0.971	**0.843**	0.975	**0.854**	0.976	**0.905**	0.964	0.934
情報源フィルタ	情報源ブログ記事	0.607	0.498	0.628	0.500	0.614	0.520	0.597	0.367
	非情報源ブログ記事	0.773	0.725	0.801	0.721	0.786	0.744	0.795	0.633
拡大情報源フィルタ	拡大情報源ブログ記事	0.709	0.556	0.739	0.553	0.738	0.557	0.720	0.505
	非拡大情報源ブログ記事	0.709	0.661	0.731	0.643	0.720	0.662	0.724	0.495
感想フィルタ	感想ブログ記事	0.946	**0.749**	0.954	**0.769**	0.946	**0.817**	0.945	0.905
	非感想ブログ記事	0.132	0.227	0.180	0.220	0.203	0.238	0.102	0.095
平均		0.628	0.554	0.652	0.555	0.650	0.577	0.619	0.500

もの（IG&RBF）の 4 種類の手法による分類精度を比較対象として示す．なお属性値には，ブログ記事本文における単語の出現数（TF）を用いた．

3.2　学習対象と同一話題に対する分類性能

表 3，表 5 のそれぞれのデータセットに対して，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（F-measure）を 2.2 節で示した 4 つのフィルタに対して計測した結果を**表 6**，**表 7** に示す．表 6 および表 7 の各フィルタに対する値は，5 つの話題での計測結果の平均値である．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 6，表 7 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．すべてのクラスの F 値をアルゴリズムごとに平均をとったものを表中の平均に示す．ここでの F 値は，あるクラスに分類された文書集合中の正解文書集合の割合を適合率（P），あるクラスに分類されるべき文書集合中の正解文書集合の割合を再現率（R）としたときの，調和平均をとったものであり，式 (3) で表される．

$$F\text{-}measure = \frac{2 \times P \times R}{P \times R} \tag{3}$$

たとえば，スパムフィルタでのスパムブログ記事クラスに対する分類精度（F 値）は，0.177 となっているが，これは，分類対象のブログ集合全体を S，S のうちでスパムブログ記事として分類されるべき記事集合を S_{opt}，S_{opt} に対してスパムフィルタが正しく分類した記事集合を S_p，スパムフィルタが誤って S_{opt} 以外のブログ記事に対してスパムと判定してしまった記事集合を S_{fp} とすると，

表 8　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本手法と Ni らの手法の分類性能（F 値）の比較
Table 8　Performance of classification filters applied to the learned events (F-measure).

フィルタ	クラス	提案手法（Boosting C4.5）	提案手法（SVM）	Ni(linear)	Ni(RBF)	Ni(IG&linear)	Ni(IG&RBF)	ランダム
スパムフィルタ	スパムブログ記事	0.360	0.489	0.497	0.365	0.489	0.365	0.050
	非スパムブログ記事	0.978	0.981	0.989	0.987	0.988	0.987	0.950
情報源フィルタ	情報源ブログ記事	0.718	0.708	0.816	**0.400**	0.813	**0.398**	0.432
	非情報源ブログ記事	0.777	0.794	0.859	0.642	0.865	0.645	0.568
拡大情報源フィルタ	拡大情報源ブログ記事	0.836	0.835	0.783	0.708	0.780	0.638	0.632
	非拡大情報源ブログ記事	0.724	0.684	0.663	**0.143**	0.654	**0.136**	0.368
感想フィルタ	感想ブログ記事	0.958	0.942	0.955	0.958	0.953	0.958	0.921
	非感想ブログ記事	0.298	0.213	0.199	**0**	0.173	**0**	0.079
平均		0.706	0.706	0.720	0.525	0.714	0.516	0.500

表 9　Google ブログ検索を用いて収集した集合における同一話題でのフィルタの本手法と Ni らの手法の分類性能（F 値）の比較
Table 9　Performance of classification filters on Google-based crawled articles applied to the learned events (F-measure).

フィルタ	クラス	提案手法（Boosting C4.5）	提案手法（SVM）	Ni(linear)	Ni(RBF)	Ni(IG&linear)	Ni(IG&RBF)	ランダム
スパムフィルタ	スパムブログ記事	0.217	0.104	0.379	0	0.378	**0**	0.066
	非スパムブログ記事	0.976	0.964	0.977	0.965	0.978	0.965	0.934
情報源フィルタ	情報源ブログ記事	0.614	0.597	0.651	**0.270**	0.697	**0.269**	0.367
	非情報源ブログ記事	0.786	0.795	0.797	0.672	0.830	0.671	0.633
拡大情報源フィルタ	拡大情報源ブログ記事	0.738	0.720	0.667	**0.306**	0.702	**0.275**	0.505
	非拡大情報源ブログ記事	0.720	0.724	0.669	0.577	0.737	0.573	0.495
感想フィルタ	感想ブログ記事	0.946	0.945	0.939	0.949	0.949	0.949	0.905
	非感想ブログ記事	0.203	0.102	0.251	0	0.265	0	0.095
平均		0.650	0.619	0.666	**0.467**	0.692	**0.463**	0.500

$P = |S_p| / |S_{opt} + S_{fp}|$

$R = |S_p| / |S_{opt}|$

であり，この P と R の値から計算した F 値（の 5 つの話題での平均値）が 0.177 であったことを示している．

表 6 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見ると，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．この傾向は，表 7 の Google ブログ検索を用いて収集した集合における同一話題でのフィルタの分類性能における，情報源フィルタや拡大情報源フィルタにおける結果でも見られる．感想フィルタの非感想ブログ記事への分類精度は，SVM よりも，Boosting C4.5 や Boosting NB が 0.08 程度良い結果を示している．この傾向は，表 7 の Google ブログ検索を用いて収集した集合における同一話題でのフィルタの分類性能における，感想フィルタにおける結果でも見られる．

表 3，表 5 のそれぞれのデータセットに対して，同様の手順で，Ni らの手法で作成したフィルタの，学習に使用した話題と同一の話題における分類精度を 10 分割交差検定法で評価した．Ni らの手法の分類精度と，表 6，表 7 中から本提案手法において比較的良い分類精度を示した Bagging C4.5 と SVM を学習アルゴリズムとして用いたときの分類精度をまとめたものを**表 8**，**表 9** に示す．

表 8，表 9 より，本提案手法と Ni らの手法の最も良い分類精度のものを比較した場合，

表 10　トラックバックを用いて収集した集合における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）
Table 10　Performance of classification filters applied to Blogs for different events (F-measure).

フィルタ	クラス	C4.5	NB	Bagging C4.5	Bagging NB	Boosting C4.5	Boosting NB	SVM	ランダム
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	0.050
	非スパムブログ記事	0.957	0.962	0.975	0.965	0.965	0.963	0.962	0.950
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	0.432
	非情報源ブログ記事	0.653	0.747	0.716	0.753	0.730	0.737	0.737	0.568
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	0.632
	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	0.368
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**	0.921
	非感想ブログ記事	0.136	0.188	0.130	0.169	0.159	0.175	0.230	0.079
平均		0.609	0.606	0.618	0.609	0.630	0.611	0.655	0.500

表 11　Google ブログ検索を用いて収集した集合における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）
Table 11　Performance of classification filters on Google-based crawled articles applied to Blogs for different events (F-measure).

フィルタ	クラス	C4.5	NB	Bagging C4.5	Bagging NB	Boosting C4.5	Boosting NB	SVM	ランダム
スパムフィルタ	スパムブログ記事	**0.036**	0.124	**0.038**	0.126	0.086	0.131	**0.036**	0.066
	非スパムブログ記事	0.956	**0.837**	0.955	**0.845**	0.955	**0.904**	0.956	0.934
情報源フィルタ	情報源ブログ記事	0.470	0.515	0.500	0.518	0.510	0.512	0.500	0.367
	非情報源ブログ記事	0.663	0.729	0.719	0.731	0.706	0.729	0.721	0.633
拡大情報源フィルタ	拡大情報源ブログ記事	0.597	0.527	0.639	0.522	0.638	0.550	0.610	0.505
	非拡大情報源ブログ記事	0.613	0.649	0.621	0.634	0.641	0.657	0.644	0.495
感想フィルタ	感想ブログ記事	0.923	**0.725**	0.937	**0.728**	0.920	**0.779**	0.931	0.905
	非感想ブログ記事	**0.048**	0.186	**0.053**	0.184	0.121	0.187	**0.077**	0.095
平均		0.538	0.536	0.558	0.536	0.572	0.556	0.559	0.500

学習対象と同一話題に対する分類の精度では，フィルタによって異なるものの，トラックバックを用いて収集した集合では平均において 0.014，Google ブログ検索を用いて収集した集合では平均において 0.042，本提案手法のものが劣るという結果が得られた．

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 3，表 5 のそれぞれのデータセットに対して，3.2 節で用いたものと同様の学習アルゴリズムである，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て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 2.2 節で示した 4 つのフィルタごとに平均をとり，まとめたものを**表 10**，**表 11** 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．すべてのクラスの F 値をアルゴリズムごとに平均をとったものを表中の平均に示す．なお，表 10，表 11 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．

また，同様の手順で，Ni らの手法で作成したフィルタの，学習に使用した話題とは異なる話題を分類した際の分類精度を評価した．Ni らの手法の分類精度と，表 10，表 11 中から本提案手法において比較的良い分類精度を示した Bagging C4.5 と SVM を学習アルゴリズムとして用いたときの分類精度をまとめたものを**表 12**，**表 13** 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．

表 6，表 10 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

表 12　トラックバックを用いて収集した集合における学習対象とは異なる話題に対するフィルタの本手法と Ni らの手法の分類性能（F 値）の比較
Table 12　Performance of classification filters applied to Blogs for different events (F-measure).

フィルタ	クラス	提案手法（Boosting C4.5）	提案手法（SVM）	Ni(linear)	Ni(RBF)	Ni(IG&linear)	Ni(IG&RBF)	ランダム
スパムフィルタ	スパムブログ記事	0.135	0.268	**0.040**	0.117	**0.018**	0.105	0.050
	非スパムブログ記事	0.965	0.962	0.976	0.976	0.975	0.977	0.950
情報源フィルタ	情報源ブログ記事	0.647	0.689	**0.132**	**0.141**	**0.219**	**0.148**	0.432
	非情報源ブログ記事	0.730	0.737	0.732	0.570	0.729	0.673	0.568
拡大情報源フィルタ	拡大情報源ブログ記事	0.804	0.780	**0.285**	**0.618**	**0.347**	**0.615**	0.632
	非拡大情報源ブログ記事	0.650	0.659	0.616	**0.099**	0.552	**0.098**	0.368
感想フィルタ	感想ブログ記事	0.950	**0.920**	0.959	0.958	0.956	0.958	0.921
	非感想ブログ記事	0.159	0.230	**0.046**	**0.011**	0.084	**0**	0.079
平均		0.630	0.655	0.467	0.436	0.485	0.447	0.500

表 13　Google ブログ検索を用いて収集した集合における学習対象とは異なる話題に対するフィルタの本手法と Ni らの手法の分類性能（F 値）の比較
Table 13　Performance of classification filters on Google-based crawled articles applied to Blogs for different events (F-measure).

フィルタ	クラス	提案手法（Boosting C4.5）	提案手法（SVM）	Ni(linear)	Ni(RBF)	Ni(IG&linear)	Ni(IG&RBF)	ランダム
スパムフィルタ	スパムブログ記事	0.086	**0.036**	0.051	0	0.096	0	0.066
	非スパムブログ記事	0.955	0.956	0.963	0.965	0.962	0.965	0.934
情報源フィルタ	情報源ブログ記事	0.510	0.500	0.298	0.234	0.291	0.090	0.367
	非情報源ブログ記事	0.706	0.721	0.705	0.451	0.631	0.595	0.633
拡大情報源フィルタ	拡大情報源ブログ記事	0.638	0.610	0.474	0.135	0.412	0.162	0.505
	非拡大情報源ブログ記事	0.641	0.644	0.552	0.514	0.592	0.479	0.495
感想フィルタ	感想ブログ記事	0.920	0.931	0.936	0.949	0.945	0.951	0.905
	非感想ブログ記事	0.121	0.077	0.062	0	0.081	0	0.095
平均		0.572	0.559	0.505	0.406	0.501	0.405	0.500

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，学習対象とは異なる話題に対しても本論文で提案する属性は適用可能である．

表 12，表 13 より，学習に使用した話題とは異なる話題に適用した場合には，本提案手法と Ni らの手法の最も良い分類精度のものを比較した場合，感想フィルタの感想ブログ記事への分類精度は本提案手法がわずかに劣るものの，トラックバックを用いて収集した集合においては平均で 0.170，Google ブログ検索を用いて収集した集合においては平均で 0.067 ほど本提案手法の分類精度が良い結果が得られ，話題に依存しない属性を用いているため，Ni 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．

表 8，表 9 で示した，学習に使用した話題と同一の話題における分類精度と比較する．本提案手法の学習アルゴリズムとして SVM を用いた場合，学習対象とは異なる話題での分類精度がトラックバックを用いて収集した集合においては平均で 0.051，Google ブログ検索を用いて収集した集合においては平均で 0.060 ほど落ちているが，Ni らの手法（IG&linear）の場合，トラックバックを用いて収集した集合においては平均で 0.229，Google ブログ検索を用いて収集した集合においては平均で 0.191 ほど分類精度が落ちており，Ni らの手法に比べ，本提案手法がブログ記事で扱われる話題に依存しにくいことを確認した．

3.4　コミュニティの形成度と分類性能

表 14 に，ブログ記事の収集方法による，収集記事数の違い，およびそれら両方の収集方法で同一の記事が得られた件数を示す．Google ブログ検索による方法では収集ブログ記事数が多くなっているが，収集された記事の多くは，トラックバックにより収集された記事とは異なっている．

ブログ記事の収集方法に対する各記事の傾向の違いについて，分類に用いた属性のうちで

表 15　収集方法によるブログ記事集合の違い
Table 15　Link-related attributes of Blog articles crawled by different methods.

収集方法	トラックバック	相互トラックバック	リンク	相互リンク	集合内リンク	共起リンク	ニュースサイト	アフィリエイト	総記事数
トラックバック	4064 (7.416)	1498 (2.916)	3511 (6.407)	10 (0.018)	169 (0.308)	417 (0.761)	612 (1.117)	506 (0.923)	548
Google ブログ検索	— (—)	— (—)	5016 (3.024)	0 (0)	15 (0.009)	545 (0.329)	382 (0.230)	602 (0.363)	1659

（括弧内は 1 記事あたりの数）

表 14　収集方法別の収集件数と同一記事数
Table 14　Same Blog articles in two datasets with different crawling methods.

話題	収集件数		同一記事数
	トラックバック	Google ブログ検索	
宮崎県知事選でそのまんま東氏当選	104	330	13
自民党総裁選で福田新総裁誕生	104	225	9
ZARD の坂井泉水さん死去	195	462	31
光市母子殺害事件 元少年に死刑判決	76	266	4
相棒-劇場版-公開	69	416	24

リンクに関連した統計情報をまとめたものを，**表 15** に示す．表 15 では，収集方法ごとに収集された記事数が異なるため，括弧内に 1 記事あたりの平均数を示している．表 15 より，Google ブログ検索により収集したブログ記事集合に比べて，トラックバックにより収集したブログ記事集合のほうが，特に集合内リンクおよび共起リンクが多く含まれ，相互に密接に関連したブログ記事集合を比較的多く収集できていたことが分かる．

表 12 と表 13 での結果の比較から，分類学習データと異なる話題に分類器を適用した場合の分類性能について，Ni らの手法では，データの収集方法の違いによる分類性能の差があまり明確でないが，本論文での提案手法では，トラックバックによって収集されたブログ記事集合に対する結果が，Google ブログ検索による記事収集の場合と比較して，明確に高い分類性能を示していることが分かる．また，表 10 と表 11 の比較から，本手法におけるこの特長は，用いる分類学習アルゴリズムの種類が違っていても，同様の傾向を示すことが分かる．すなわち，本手法では，ブログ記事が持つコミュニティとしての性質をうまく保存した状態で収集できた場合に，その性質をうまく利用して分類学習を行えていることが分かる．

3.5 議　　論

表 10 での分類性能をフィルタ別にみると，スパムフィルタについては，平均で 0.14 程度の精度でスパムブログ記事を分類している．この精度を下げていた要因として，実験対象としたデータセットにおけるスパムの少なさがあげられる．特に，表 4 にあるように，「宮崎県知事選でそのまんま東氏当選」「自民党総裁選で福田新総裁誕生」の話題では，スパムブログ記事と定義できるブログ記事はそれぞれ 1 記事しか存在しなかったため，学習事例が極端に少量となり学習アルゴリズムが効果的に働かなかったと考えられる．スパムブログ記事が 1 記事しか存在しなかった前述の 2 話題以外の 3 話題を対象としスパムブログ記事への分類精度を F 値の平均で見た場合，最も良いもので SVM の 0.512，最も悪いもので Boosting NB の 0.241 であった．すべての学習アルゴリズムの分類性能の平均では，0.402 であった．学習データ内での分類結果が極端に偏っていて，有効な分類事例がごく少数しかない場合への対処は今後の課題である．

4. 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．1 つが，ブログ記事の内容の時系列・空間解析である．Gruhl らは，ブログ空間における情報の広がりをモデル化する[17]．Glance らは，ブログ空間において様々に隆起しているトレンドを発見する手法を提案している[18]．Mei らは，時系列と空間における話題の移り変わりをとらえるための確率的アプローチを提案している[1]．もう 1 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．Leshed らは，ブログ記事のテキストからブログ著者の感情を読み取るシステムを提供している[19]．Mishne は，投稿されたブログ記事において，confused や sad や excited や happy など，37 の感情ごとにブログ記事を分類する実験を行っている[20]．Ni らは，ブログ記事を，専門技術について書かれているような Informative article とブログ著者の個人的な感情が書かれているような Affective article という 2 つの分類に分類する手法を提案している[16]．平野らは，

日本語圏ブログに対して,「エンターテイメント」「スポーツ」「生活」などの一般的なカテゴリへのリアルタイム自動分類を行っている[21].

これらの関連研究[16),19)–21)]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るもの[16)]，出現頻度の上位 N 単語までを用いるもの[19),20)]，ある閾値以上出現した単語のみを用いるもの[21)]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．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対象とし，ブログ記事の分類を行う．そのため,「エンターテイメント」「スポーツ」「生活」などの一般的なカテゴリではなく，ブログ記事の閲覧者のニーズに特化した分類クラスへの分類を行うことを可能としている．

5. おわりに

本論文では，ブログ特有のトラックバックなどによるブログ記事相互の関係性を考慮して収集したブログ記事集合に対し，そこで形成されたコミュニティの特性および品詞やリンクなどの記事内の話題に影響されにくい統計情報を分類学習時の属性として用いる手法を提案した．

分類学習のアルゴリズムとしては，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た．ブログ記事分類問題の例として，ブログ記事を「スパムフィルタ」「情報源フィルタ」「拡大情報源フィルタ」「感想フィルタ」という 4 つのフィルタを用いて,(1) 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,(2) 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,(3) 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,(4) 感想フィルタでは感想ブログ記事と非感想ブログ記事，の分類クラスに分類する際の分類性能を評価した．ブログ記事集合を，記事集合中のので形成されたコミュニティの保存度が異なる 2 種類の方法で収集し，学習データとは異なるトピックに対する分類性能の評価を行ったところ，特にコミュニティの保存度が高い方法で収集したブログ記事集合に対して，他の手法と比較して本手法では学習データが扱う話題に強く依存せず，高い分類性能を発揮できることを示した．統計情報に基づいた属性と機械学習を用いているため，本論文で示したブログ記事分類問題とは異なる分類目的にも本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合せによる分類，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．

本分類手法は，トラックバックを用いて収集した集合に対しての分類精度は他の手法と比較して十分な性能が得られているが，Google ブログ検索を用いて収集した集合に対しては精度の低下が見られ，Boosting NB 以外の学習アルゴリズムによる分類ではランダムな分類に精度が劣る場合が見られた．本分類手法は，トラックバックに基づくブログ記事の収集時には性能を発揮するが，Google ブログ検索などによって収集されたブログ記事集合では必ずしも他の手法と比較した明確な利点があるとはいえない．トラックバックコミュニティの広範囲なブログ記事からの効果的な収集方法の実現，および，トラックバックコミュニティを十分に形成していないブログ記事集合に対する効果的な分類手法の実現については，今後の検討課題である．

謝辞　本論文の洗練にあたり，適切で的を射た指摘を数多くいただいた査読者の方々に心から感謝する．

参 考 文 献

1) Mei, Q., Liu, C., Su, H. and Zhai, C.: A probabilistic approach to spatiotemporal theme pattern mining on weblogs, *WWW '06: Proc. 15th international conference on World Wide Web*, New York, NY, USA, pp.533–542, ACM (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 (2006).
3) 戸田智子，黒田晋矢，福田直樹，石川　博：ブログにおける多視点からのトピック抽出手法の提案，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，B4-2 (2008).
4) Google ブログ検索．http://blogsearch.google.com/
5) 鎌田基之，横山昌平，福田直樹，石川　博：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，第 1 回データ工学と情報マネジメントに関するフォーラム，i1-31 (2009).
6) 石田和成：スパムブログの推定と抽出，日本データベース学会論文誌（DBSJ Letters），Vol.6, No.4, pp.37–40 (2008).
7) 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).

8) Quinlan, J.R.: *C4.5: programs for machine learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993).
9) Langley, P., Iba, W. and Thompson, K.: An analysis of Bayesian classifiers, *Proc. 10th National Conference on Artificial Intelligence*, Vol.228, pp.223–228, AAAI Press, Menlo Park, USA (1992).
10) Cortes, C. and Vapnik, V.: Support Vector Networks, *Machine Learning*, pp.273–297 (1995).
11) Breiman, L. and Breiman, L.: Bagging predictors, *Machine Learning*, pp.123–140 (1996).
12) Freund, Y. and Schapire, R.E.: Experiments with a New Boosting Algorithm, *Proc. 13th International Conference on Machine Learning*, pp.148–156, Morgan Kaufmann (1996).
13) Witten, I.H. and Frank, E.: *Data Mining: Practical machine learning tools and techniques*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005).
14) Fan, R.-E., Chen, P.-H. and Lin, C.-J.: Working Set Selection Using Second Order Information for Training Support Vector Machines, *J. Mach. Learn. Res.*, Vol.6, pp.1889–1918 (2005).
15) Chang, C.-C. and Lin, C.-J.: *LIBSVM: A library for support vector machines* (2001). Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsvm
16) Ni, X., Xue, G.-R., Ling, X., Yu, Y. and Yang, Q.: Exploring in the weblog space by detecting informative and affective articles, *WWW '07: Proc. 16th international conference on World Wide Web*, New York, NY, USA, pp.281–290, ACM (2007).
17) Gruhl, D., Guha, R., Liben-Nowell, D. and Tomkins, A.: Information diffusion through blogspace, *WWW '04: Proc. 13th international conference on World Wide Web*, New York, NY, USA, pp.491–501, ACM (2004).
18) Glance, N., Hurst, M. and Tomokiyo, T.: BlogPulse: Automated Trend Discovery for Weblogs, *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, Vol.2004 (2004).
19) Leshed, G. and Kaye, J.J.: Understanding how bloggers feel: recognizing affect in blog posts, *CHI '06: CHI '06 extended abstracts on Human factors in computing systems*, New York, NY, USA, ACM, pp.1019–1024 (2006).
20) Mishne, G.: Experiments with mood classification in blog posts, *Style2005—1st Workshop on Stylistic Analysis of Text for Information Access, at SIGIR 2005* (2005).
21) 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).

(平成 20 年 12 月 20 日受付)
(平成 21 年 4 月 15 日採録)

(担当編集委員　岩山 真)

鎌田　基之

2007 年静岡大学情報学部情報科学科卒業．同年同大学大学院情報学研究科修士課程入学．2009 年同大学院修了．同年京セラ株式会社入社．現在に至る．修士（情報学）．Web，ブログのクローリング技術やブログからのマイニングに興味を持つ．日本データベース学会会員．

福田　直樹（正会員）

1997 年名古屋工業大学工学部知能情報システム学科卒業．1999 年同大学大学院工学研究科電気情報工学専攻博士前期課程修了．2002 年同大学院博士後期課程修了．同年静岡大学情報学部情報科学科助手．2007 年より同助教．現在に至る．博士（工学）．セマンティック Web，Web サービス，電子商取引，およびエージェントシステムの実装技術に関する研究に従事．IEEE-CS，ACM，電子情報通信学会，人工知能学会，ソフトウェア科学会，情報システム学会各会員．

横山　昌平（正会員）

静岡大学情報学部助教．産業技術総合研究所を経て 2008 年より現職．2006 年東京都立大学大学院工学研究科修了，博士（工学）．データベース技術の研究開発に従事．電子情報通信学会，日本データベース学会各会員．情報処理学会論文誌（データベース）幹事補佐，電子情報通信学会データ工学研究専門委員会幹事補佐．

石川　　博（正会員）

静岡大学情報学部情報科学科教授．東京大学理学部情報科学科卒業．東京都立大学を経て 2006 年より現職．東京大学博士（理学）．著書に『次世代データベースとデータマイニング— DB&DM の基礎と Web・XML・P2P への適用』（CQ 出版社),『JavaScript によるアルゴリズムデザイン— オブジェクト指向から DB・Web・マイニングまで』（培風館),『データベース』（森北出版）等．国際論文誌 ACM TODS，IEEE TKDE，国際学会 VLDB，IEEE ICDE 等学術論文多数，1994 年情報処理学会坂井記念特別賞，1997 年科学技術庁長官賞（研究功績者）受賞．情報処理学会データベースシステム研究会主査，情報処理学会（データベース）共同編集委員長，International Journal Very Large Data Bases Editorial Board，日本データベース学会理事歴任．ACM，IEEE，電子情報通信学会各会員．