社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2005－NL－170 (4)
2005／11／21

日本語圏ブログの自動分類

平野耕一† 古林紀哉‡ 高橋淳一‡

† NRI Pacific, Inc. 1400 Fashion Island Blvd. Suite 1010. San Mateo, CA 94403, U.S.A.
‡ 株式会社野村総合研究所. 情報技術本部 〒240-0005 横浜市保土ヶ谷区神戸町１３４
E-mail: †koichi@nri.com, ‡n5-kobayashi@nri.co.jp, ‡j-takahashi@nri.co.jp

あらまし: 消費者の個別のニーズや文脈を検出することは、ネットビジネス企業にとってだけでなく、消費者自身の生活者の質向上のためにも非常に重要である。最近のブログの流行によって、消費者の日常生活についての情報や日々考えていることを低コストかつリアルタイムに取得することが可能になった。本論文では、ナイーブベイズ法に基づく多重トピック分類などの自然言語処理技術を、日本語圏ブログのリアルタイム分類とトピック定量化に適用した最初の試みを報告する。本論文で提案した方法によって、ブログエントリのリアルタイム多重トピック自動分類と、ブログ圏における多重トピックの強度の定量化が可能であることを実証した。

キーワード：ブログ、文書分類、ナイーブベイズ、多重クラス分類

Text Categorization for Japanese Blog Entries

Koichi HIRANO†, Noriya KOBAYASHI‡, Junichi TAKAHASHI‡

† NRI Pacific, Inc. 1400 Fashion Island Blvd. Suite 1010. San Mateo, CA 94403, CA. U.S.A.
‡ Nomura Research Institute, Ltd. 134 Godo-cho, Hodogaya-ku, Yokohama 240-0005, Japan
E-mail: †koichi@nri.com, ‡n5-kobayashi@nri.co.jp, ‡j-takahashi@nri.co.jp

Abstract: Identifying individual's needs and context is of paramount importance not only from the net business player's perspective but also for improving consumers' quality of life. The recent blogging boom provides consumers' everyday life information and thinking accessible at a low cost in real time. This paper discusses the first attempt to apply some NLP techniques, such as multi-class document classification build upon the naïve bayes method, to a real-time classification and quantification of their topics in the Japanese blogosphere. The results show that the proposed method is capable of classifying blog entries in real-time and is also capable of quantifying the intensity of multi-topics in the blogosphere.

Keywords: blog, document classification, multi-class categorization, naïve bayes

1 はじめに

本論文では、大量に生成されつつある日本語圏ブログエントリについてディレクトリ型リアルタイム自動分類の一つの試みを報告する。近年、日本語圏においてもブログの流行は止まるところを知らない。「ブログサービスサイト比較 2004」[1]に報告されているインターネット調査（調査日程 2004/8/6～12, n=14,542）によるとブログ所有者はインターネットユーザ全体の 5.9%、今後のブログ意向者はインターネットユーザ全体の 18.2%であった。また総務省の報道資料[2]によると、2005 年 3 月末時点の国内ブログ利用者数は延べ約 335 万人とされている。既に、人口の無視できない部分がインターネット上で、自ら考えていること、感じたことを発信する状況になっている。多数の生活者の生の声がブログの形で蓄積しつつあり、梅田はこの様相を「総表現社会」[3]と称した。

「総表現社会」の一つの特異性は、Wired Magazine 編集長 Chris Anderson が名づけた "The Long Tail" 分布[4]の長大な尻尾部分についての情報を、インターネットを経由してほぼ無視できるコストで取得可能になった点にある。"The Long Tail" とは、Zipf の法則やべき乗則に従うような頻度分布が長大である統計的性質のカジュアルな呼称で、ネット社会でのいくつかの事例が報告されている。ビジネス的には、ニッチな個（あるいはニーズ）もかき集めるとその潜在的な市場の大きさは、主流の個の市場を超える、ということに "The Long Tail" の意義があり、ニーズの個別性をいかに低

コストで検知しそれに対処するかが一般的課題となる。

個別的なニーズの内容検知という課題は難渋を極める。これに対する一つの擬似的解決手段は、ブログエントリの多重トピック自動分類が考えられる。すなわち、エントリ本文が「*今日は朝から家で過す。私は、家事に精を出し、子供は派手に遊んでいる。しかし、一日中殆どけんか無しで良く遊べるなぁ。今日の昼は、豚骨と骨付き鶏でとったスープを使っておじや。具はキャベツと玉ねぎ。子供はがつがつ食べている。*」というブログエントリは、例えば「出産・育児」と「料理・レシピ」といった二つのトピックに分類しうる。この例のように、一つのブログエントリが2つ以上のトピックを持つこと、あるいは、生活者が複数のトピックを持つことは一般的である。例えば分類数を 100 とする分類体系において、ある生活者のある日のトピックを3つ特定することは、約 100 万ある可能的多重トピックの組み合わせのうち1つに絞り込んだことを意味する。かなり個別的に生活者の声を捉えることが可能になる。例えば、生活者のある2つのトピックをターゲットとしたサービスを持つ企業が、3つ目のトピックとしてどういうものがあるのかを発見できることは、マーケットプロモーション上非常に有用な情報となりうる。

ブログの分類、ということに関連した最近の動きとして、フォークソノミー(Folksonomy)あるいはソーシャル・タギング(Social Tagging)と呼ばれる仕組みがブログ利用者の中で流行しており、例としてオンラインブックマークサービスの“del.icio.us”[5]や“はてなブックマーク”[6]、オンラインアルバムサービスの “flickr”[7]があげられる。これらは、ブログコミュニティ参加者の「自己申告」による分類と見なすことができる。これら「自己申告」の情報を参照し様々な関係情報を引き出す研究も見られる[8]。一方で自己申告に頼らない第三者的な分類も補完的な情報として価値がある可能性があり、本研究ではこの観点にたち、日本語圏ブログのリアルタイム自動分類を試みる。

残念ながら、日本語圏ブログのリアルタイム自動分類には多くの運用上、技術上の課題がある。いくつか挙げると：

(i) より意味ある分類体系をいかに作成するか（解像度はある程度高くないといけない）、
(ii) 分類毎の訓練データをどう準備するか（分類数が多ければ大変な手作業が発生する、また分類中の話題も日々進化している）、
(iii) 学習速度、
(iv) 多トピック自動分類の精度ならびに評価方法、
(v) 自動分類の速度（2005 年 10 月 10 日現在、日本語のブログエントリは秒間2個以上生成されていることが、日本の代表的 ping サイト ping.bloggers.jp[9]のデータより推定できる）、
(vi) ブログエントリのタイトル及び本文の抽出。

などがある。本論文では、以上にあげた個々の問題に対する最適解を追求するというよりも、まず、日本語圏ブログエントリのリアルタイム自動分類により何が見えてくるか、に主眼を置いた一つの速効的取り組みについて論じる。

自動分類は、前述の問題意識から、生活者の視点により近いものを得ることを意図した。続く 2 章「ブログエントリの自動分類」では、まず、分類体系と訓練データについて、次に多重トピック分類の方法と速度問題について、3 章ではある期間の日本語圏ブログの分類実験結果と考察を述べる。4 章では今後の課題と本研究の拡張の方向性を述べる。

2 ブログエントリの自動分類手法

2.1 分類体系と初期訓練データ

日本語圏で利用可能な文書の分類体系には、国際十進分類法(UDC)、日本十進分類法(NDC)、国立国会図書館分類表(NDLC)など主に図書を分類するための体系と、Yahoo!カテゴリや goo カテゴリなど Web ページへのナビゲートを目的とした体系がある。前者は学術的な知識の分類を目指しており学術系出版系の多くの団体から統一的に参照され見直しは 10 年くらいの長期間隔で行われる。一方後者は、ポータルサイト独自にインターネット利用者の興味を中心に編集されるため、体系見直しの間隔も短く各ポータルサイトで様々な構成となっている。

本研究では、1 章で述べた文脈から、複数のポータルサイトのカテゴリ体系を参考にして表 1 に示すような独自の2階層分類体系を生活者の視点から作成、採用した。

初期訓練データもより生活者の視点に立ったものに必要がある。ネットコミュニティにより形成された主観的文書を採用した自動分類の優れた試みとして、阿部らの研究がある[10]。阿部らは、権威ある機関が策定した分類基準による分類を唯一無二の正解とした分類でなく、よりユーザの主観性に近い分類が行える可能性を示唆している。本研究でも、同様の立場をとり、Yahoo!掲示板[11]の投稿を初期訓練データとして用いた。Yahoo!掲示板での分類体系は、主催者である Yahoo!Japan が採用・決定したものであるが、ある投稿内容がどの分類に属するかは、掲示板投稿者の決定事項である。トピック（その分類体系）にそぐわない投稿は、掲示板コミュニティによって歓迎されず、それがコミュニティの自律的な制御機能を果たしている。

我々の用いる分類体系そのものは、筆者らが任意に策定したものではあるが、初期訓練データに掲示板投稿記事を用いることにより、分類基準はコミュニティが形成したコンセン

サスに依拠していることになる。これにより、より生活者の視点に近い分類が期待できる。

2.2 ブログ自動分類器の構成

2.2.1 初期訓練データ取得機構

初期訓練データ取得機構は、Yahoo!掲示板より、あらかじめ指定した複数の「トピ」と呼ばれる投稿をクロールし、本文を取得してデータベースに蓄積する。

初期訓練データと分類体系の対応に関しては後述する。

2.2.2 自動分類機構・学習機構

文書自動分類の方法としては、ナイーブベイズ法、SVM法、ブースティング法、決定木による方法など多くのものが提案されている[12]。さらに、多重トピック自動分類については、対象文書が分類に属するかどうかの2クラス分類器を次元数分組み合わせた方法、Parametric Mixture Model、最大マージンラベリング法などが提案されている[13]。本研究では、以下の理由でオープンソースのスパムフィルタで実績のあるGary Robinsonが提案するベイズ法[14]による2クラス分類器の組み合わせによる多重トピック自動分類を行った。

第一に、一般にベイズ法は分類結果の解析が比較的容易な手法である。巨大掲示板の記事群による初期訓練データから特定の分類体系に従った分類の実現可能性に関しては、個々の分類に対し、分類結果が既知である掲示板投稿記事のカテゴリ帰属確率の分析により論じることができる。

第二に、2クラス分類器の組み合わせによる自動分類結果は、自然に多重トピック自動分類になる。一般に多重トピック分類された訓練データを大量に入手することは困難であり、本研究で用いた初期訓練データの記事も意味上多重トピックを持っていたとしても、単一トピックの情報しか持っていない。しかしながら、2クラス分類器の組み合わせは結果として、多重トピックの結果を出力することができる。

以下、個々の分類のことをカテゴリと呼び、本研究で用いた手法を説明する。

訓練データに含まれるある単語w_kが、あるカテゴリc_iに現れる頻度をF_{ik}、c_i以外のカテゴリに出現する頻度を$F_{\bar{i}k}$としたとき、単語w_kを含む文章がカテゴリc_iに含まれる確率を

$$P(w_k \mid c_i) = \frac{F_{ik}}{F_{ik} + F_{\bar{i}k}} \cdots (1)$$

と定義する。

ある文章Eに含まれる単語を$\{w_1,\ldots,w_n\}$とし、c_iカテゴリ、ならびにc_i以外のカテゴリの両方において、単語は独立に生起すると仮定すると、この文章がカテゴリc_iに含まれる事後確率は、

$$P(c_i) = \prod_{k=1}^{n} P(w_k \mid c_i) \cdots (2)$$

で求められる。これが閾値αを超えたとき、文章Eはc_iに属するとした。尚、未知単語、すなわち訓練データで出現頻度0であった単語は、簡単のため、今回は無視することとした。閾値αは、スパムフィルタでの適用事例報告を参考にし、0.55とした[14]。

表1．本研究における分類体系

第一階層	第二階層	第一階層	第二階層
アニメ・ゲーム	アニメ	生活	インテリア
	ゲーム		ペット
エンターテインメント	クラブ		家庭
	スポーツ		家電
	テーマパーク		介護
	テレビ		海外、留学
	ドラマ		教育
	映画		懸賞
	演劇		資格
	音楽		受験
スポーツ	ゴルフ		住まい
	サッカー		出産・育児
	スキー		保険
	テニス	地域情報	海外情報
	マリンスポーツ		都道府県
	モータースポーツ	美容、ファッション、健康	ジョギング
	格闘技		ダイエット
	相撲		ファッション
	野球		ブランド
ニュース	芸能		健康
	最新		美容
	事件		病気
	社会	料理、グルメ	お菓子
	政治経済		お酒
	天気		スイーツ
パソコン、インターネット	Mac		ラーメン
	PDA		レストラン
	インターネット		料理
	コンピュータ	恋愛	結婚
	パソコン		出会い
	ホームページ		恋愛
	携帯	ビジネス、経済、政治	株
趣味	アート		起業・アフィリエイト
	オークション		経済
	ガーデニング		就職、転職
	ダンス		政治
	パチンコ		税金
	韓流	芸術、学術	自然科学
	競馬		人文社会科学
	語学		美術
	自動車		文学
	写真	アダルト	アダルト
	小説	その他	ボランティア
	釣り		環境
	読書		
	漫画		
	旅行		

各カテゴリに対しての訓練データの対応付けは以下のように行った。表1に示した第二階層のカテゴリに対し、このカテゴリに属すると考えた掲示板「トピ」を関連付けた。膨大に存在する「トピ」のうち、合計7385個の掲示板を選び、第二階層の91カテゴリに結びつけた。この時、一つの掲示板は必ず91カテゴリのうちの1カテゴリにのみ属するようにし、対象掲示板の投稿の本文を訓練データとした。例えば、表1に示した第一階層「スポーツ」のサブカテゴリである「テ

ニス」の訓練データには、Yahoo!掲示板カテゴリの

ホーム > スポーツ、レジャー > スポーツ > テニス[15]

に属する「トピ」と呼ばれる掲示板の複数の投稿を$c_{テニス}$の投稿として参照した。

また、確率要素である単語は、Chasen ver2.3.3 に IPADIC ver2.6.3 を組み合わせた形態素解析器[16]により、入力文章を形態素解析し、出力の中から、名詞・形容詞・動詞のみを採用した。

2.2.3 多重2クラス分類器の高速化

ここでは、膨大な数が生成されているブログエントリへの処理速度的対処について述べる。日本の代表的な更新情報 ping サイトである ping.blogger.jp には、国内のブログサイトから1日当たり少なくとも 10 万件の更新情報が送信されている。このことから、ブログエントリを自動分類する際の処理速度の目標を、1件当たり1秒に置くことができる。そして、この処理速度を持った自動分類器を多重化することで、国内ブログエントリのリアルタイム分類が達成可能となる。残念ながら筆者等の調査では、自動分類の速度に言及している報告は見当たらなかった。

あるブログエントリEが、表1に示した 91 分類それぞれに属するかどうかを判定するには、素朴な方法をとると、2値問題自動分類を 91 回繰り返さなければならない。仮に2値問題自動分類一回の演算に 0.5 秒かかるとすれば、91 分類には約 45 秒かかってしまう。この処理時間の殆どは、単語帰属の統計情報を格納したディスク装置へのランダムアクセス時間である。

本研究では分類の処理速度を高速化するために、ベクトル演算の概念を導入し、ディスク装置へのアクセス回数を分類数に依存させないことを試みた。すなわち、次元数 j の分類体系において、$\vec{c} = (c_1,...,c_j)$とすると、式(1)は、

$$P(w_k \mid \vec{c}) = \left(\frac{F_{1k}}{F_{1k} + F_{1k}},..., \frac{F_{jk}}{F_{jk} + F_{jk}} \right) \quad \cdots (3)$$

となり、式(2)は、

$$P(\vec{c}) = \left(\prod_{k=1}^{n} P(w_k \mid c_1),... \prod_{k=1}^{n} P(w_k \mid c_j) \right) \cdots (4)$$

と表せる。このことにより、ディスク装置から CPU へのデータ転送量は2値問題自動分類器の繰返し適用の場合と変わらないが、データ転送回数を分類数によらず一定に押さえることができる。

2.2.4 ブログ取得機構

取得するブログエントリの URL リストを作成し、実際その html を取得する機構である。

まず、ping.bloggers.jp に更新情報 ping を送信したブログサイトのリスト(changes.xml)を定期的に取得する。chagens.xml と呼ばれるリストには、新規のエントリ投稿があったブログの URL、ブログタイトル、ブログの説明、更新日時は含まれるが、記事本体であるブログエントリの URL は含まれない。そこで、changes.xml のリストをもとに、実際のブログのトップページURLが指す html ドキュメントを取得し、その中に記述されている RSS フィードの URL 情報を認識して、RSS フィードを取得する。さらに RSS フィード中にあるブログエントリ情報のリストから、個別のブログエントリを指す URL、タイトル等を取得する。最後に、そのURLが指すブログエントリ本体である html ドキュメントを取得し、ローカルなディスク装置に格納する。

2.2.5 エントリ切り出し機構

エントリの html ドキュメントから、記事の本文を切り出す機構である。

各エントリの html ドキュメント全体には、ブログサイトの説明、作者のプロフィール、他のエントリへのナビゲーション、広告などが含まれており、そのまま自動分類器への入力とすることは不適当である。また、RSS フィードの本文記述は先頭 100 文字程度のことが多く、全文の取得にはエントリの html ドキュメントからの切り出しが必要となる。現在のブログは書式が様々であるため、エントリの本文を正確に切り出すことは容易ではない。

本研究では、南野らの取り組み[17]と同様の手法で、エントリ切り出し機構を作成しそれを採用した。なお筆者らの、エントリ切り出し機構の成功確率は推定9割である。

2.2.6 フィードバック機構

2.1 で議論したように、最終的には生活者・ネットコミュニティの視点で自動分類をする、という観点から言うと、ある自動分類結果の正答が一意に定まることはない。言い換えると、様々な正答があってしかるべきである。本自動分類機構が出した結果に対する様々な人の意見やフィードバックを収集し、運営者である筆者らの判断をへて訓練データに反映するための機構が必要である。本研究では残念ながら、フィードバック機構の実装には至っていない。

2.2.7 マシンスペック

本研究では、Intel 社 Xeon 2.4GHz×2CPU, Memory 4GB を装着した IA32PC サーバ 2 台を使って実験を行った。サーバ 1 台をブログエントリの取得に、1 台をブログエントリの自動分類に当てた。

3 日本語圏ブログの分類実験

3.1 分類実験の対象

本実験では 2005 年 10 月 1 日午前零時から 1 週間の間に、ping.bloggers.jp に更新情報 ping を送信したブログのうち、国内ブログサービス事業者でアクティブユーザ数[18]の多い 18 社でのブログを対象とした（必ずしも上位 18 社にはなっていない）。 収集したデータには一部欠落もあるが、最終的に認識したブログサイト数は 163,417、ブログエントリ数はは 830,974 で、そのうち 805,324 個のブログエントリの分類に成功した。

なお訓練データには、Yahoo!掲示板から、7385 の掲示板を選択し、合計 5,719,008 投稿を事前に取得している。

3.2 実験結果

3.2.1 分類分布

対象期間中のブログエントリの第一階層分類毎の総文字数と頻度を図 1 、 2 に示す。

3.2.2 自動分類の精度

本研究では、複数の利用者からの分類結果に対するフィードバックを収集・解析を行っていない。筆者らによる自動分類結果のサンプル測定結果を述べると、7 割から 8 割のエントリが妥当なカテゴリに分類されていた。この成績は、同じくコミュニティが生成したデータを使った自動分類の阿部らの試み[10]とほぼ同等であった。

一つの自動分類例を示す。「*妊娠後期の妊婦が、夫婦で沖縄に飛行機で旅行にいったが、エコノミークラスの座席では大きなお腹が非常に苦しい思いをした。往路ではフライトアテンダントからは何の気遣いも得なかったが、復路ではフライトアテンダントが温かい気遣いをしてくれ、嬉しいと同時に、同じ航空会社でもこうも対応が異なるのかと驚いた。*」というような内容の妊婦によるエントリ[19]は、次の 3 トピックに分類された。「出産育児」、「旅行」、そして「介護」である。このエントリは狭い意味での介護の話題ではないが、ケアに関する文脈を持つことには違いない。介護という文脈があることを自動分類結果が教えているととらえられた。

3.2.3 自動分類の速度

本研究の実験環境においては、2.2.2 で述べた自動分類機構は、

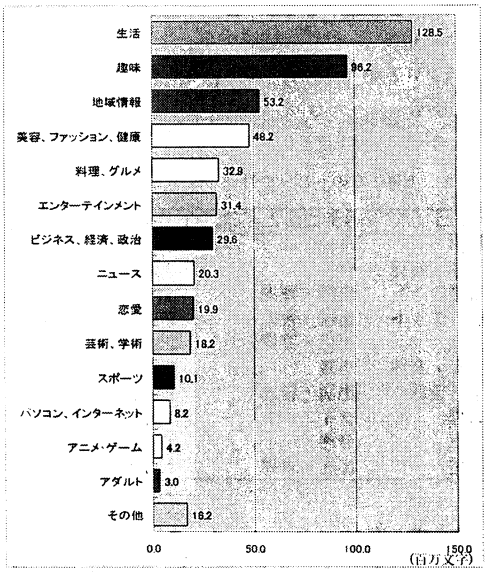

図 1 2005 年 10 月 1 日から 1 週間のブログエントリの自動分類結果（ブログエントリ本文の総文字数で集計）

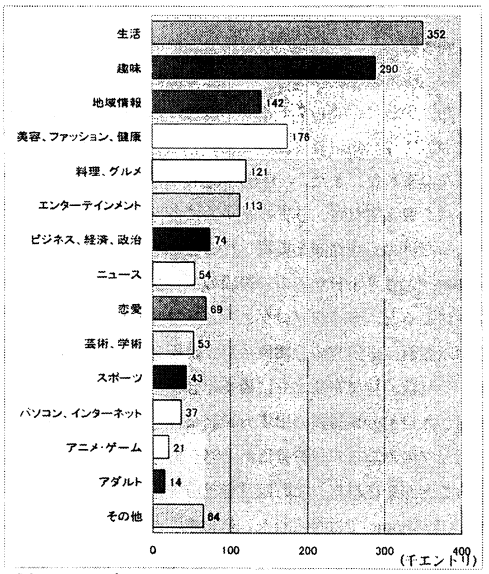

図 2 2005 年 10 月 1 日から 1 週間のブログエントリの自動分類結果（ブログエントリ数で集計）

1 ブログエントリあたり、約 2 秒で分類結果を返した。ベクトル計算を用いない場合、同一環境で約 50 秒必要だった。なお、一エントリーあたりの平均文字数は約 490 文字であった。

3.2.4 トピック多重度

一つのエントリが何重のトピックを持つかをあらわすトピック多重度は、全体の 58%のエントリが 2 以上であった。

3.2.5 トピック対の分布

あるブログエントリが持つ多重トピックのうち、２トピックの組み合わせ、すなわちトピック対に着目し、エントリ数が多いトピック対上位１０を表２に示す。

表２　上位１０のトピック対

カテゴリ１	カテゴリ２	エントリ数
読書	小説	37052
出産・育児	介護	36447
料理	出産・育児	31577
ダイエット	出産・育児	25858
病気	出産・育児	23907
出産・育児	小説	23569
海外情報	出産・育児	21581
料理	ダイエット	19943
病気	介護	19144
結婚	出産・育児	17365

3.3 まとめ

１章「はじめに」おいて、生活者視点によるブログのリアルタイム自動分類に関し幾つかの課題を提起した。このうち本報告では、(i)分類体系については、筆者らが独自に作成したものを例示した。また、(ii)分類毎の訓練データの問題については、巨大掲示板の投稿を参照することで解決する試みを示した。さらに、(iv)多トピック自動分類の精度と評価の課題については、利用者などのフィードバック機構を利用する方法を提案した。また、(v)速度問題については、数台の PC サーバを導入すれば、リアルタイムに多トピック自動分類を行えるだけの処理性能を実現したことを示した。

2005 年 10 月 1 日からの一週間のブログエントリ自動分類結果によると、多くのブログエントリは多トピックであることが示され、本研究の分類体系第二階層では「出産育児」がエントリ数、総文字数ともに最多の分類であった。本研究が採用した自動分類機構の出す分類結果は、ブログエントリのトピックのみならず、背景にある潜在的な文脈を示すこともあることが示された。これは「自己申告」をベースとしている Folksonomy では得られない情報を、本研究で提案している方法によって検出できることを強く示唆する。

4 最後に

本研究におけるブログの自動分類は三つの特徴がある。一つ目は分類体系は筆者らが決めた人為的なものであるということ、二つ目は自動分類を機械により実施しているということ、三つ目は訓練データはコミュニティが形成したものを参照している、ということにある。これらにより、日本語圏ブログでどういう話題が語られているのかや生活者が同時に持つ複数の文脈など検出することが可能となりつつある。

今後の研究課題として、利用者からのフィードバックに基づいた評価と SVM などを用いた自動分類の精度改善があげられる。また、この研究の応用の一つとして、話題の定量化があげられる。例えば、ある日のサッカーに関するブログエントリの総数と総文字数を、指標とすることが可能である。もう一つの応用例として、佐藤らが提案する話題語抽出との併用により、文脈と話題語の２軸からの話題の定量化測定が可能になる。

ブログの普及にともなう総表現社会が自然言語処理の応用分野を大きく広げていることは揺ぎ無い事実になりつつある。本研究は、その一つの証拠を付け加えたと言える。

参考文献

[1] 株式会社ビー・エム・エフティー, “ブログサービスサイト比較調査 2004”, 2004 年 11 月.

[2] 総務省報道資料, “ブログ・SNS（ソーシャルネットワーキングサイト）の現状分析および将来予測”, http://www.soumu.go.jp/s-news/2005/050517_3.html , 2005 年 5 月 17 日.

[3] 梅田, “ウェブ社会[本当の大変化]はこれから始まる”,フォーサイト, 2005 年 6 月号, pp.8-10,新潮社.

[4] Chris Anderson, “The Long Tail”, in Wired Magazine, Issue 12.10, October 2004.

[5] “del.icio.us”, http://del.icio.us/

[6] “はてなブックマーク”, http://b.hatena.ne.jp/

[7] “flickr”, http://flickr.com/

[8] 大向、松尾、松村、武田、“Community Web プラットフォーム”,第 19 会人工知能学会全国大会, 2005 年 6 月.

[9] “ping.bloggers.jp”, http://ping.bloggers.jp/

[10] 阿部, 田中, 中川, “コメントを用いた映画の自動分類”, 情報処理学会 NL 研究会報告, NL-150-16, pp.105-110, 2002 年 7 月.

[11] “Yahoo!掲示板”, http://messages.yahoo.co.jp/

[12] 永田、平, “テキスト分類－学習理論の「見本市」－”, 情報処理学会誌, Vol.42, no.1, pp32-37, 2001 年 1 月.

[13] 加沢、泉谷、平、前田, “最大マージン原理にもとづく多重トピック文書の自動分類”, 情報処理学会 NL 研究会報告, NL-163-8, pp.53-60, 2004 年 9 月.

[14] Gary Robinson, “Spam Detection”, http://radio.weblogs.com/0101454/stories/2002/09/16/spamDetection.html

[15] “Yahoo! 掲示板テニスカテゴリ”. http://messages.yahoo.co.jp/bbs?action=topics&board=1834659&sid=1834659&type=r

[16] 松本他, “形態素解析システム茶筌”, http://chasen.naist.jp/hiki/ChaSen/

[17] 南野,奥村, “なんでも RSS! －HTML 文章からの RSS Feed 自動生成”, 人工知能学会セマンティックウェブとオントロジー研究会報告, SIG-SWO-A501-03, 2005 年 7 月.

[18] “ブログファン”, http://blogfan.org/

[19] 著者不明, “ニンプ沖縄に行く”, http://taratta.at.webry.info/200509/article_9.html

[20] 古林, 平野, 高橋, “日本語圏ブログをもとにした話題指標の開発”, （準備中, 2005）.