ブログ記事の自動分類により消費者意識の側面を捉える試み

野村総合研究所
情報技術本部　上級研究員

古林 紀哉（こばやし のりや）

情報技術とネットビジネスの学際的な調査研究とシステム開発が専門。
最近の興味は、Web 2.0。博士（工学）

NRI Pacific ディビジョンマネージャ

平野 耕一（ひらの こういち）

専門分野は、計算機科学理論のビジネス適用。Ph.D.

野村総合研究所
情報技術本部　副主任テクニカルエンジニア

高橋 淳一（たかはし じゅんいち）

情報技術本部にてシステム基盤の設計および構築を担当。
最近は、IT 技術を利用した新しいビジネスの創造に従事。

要旨

最近のブログの流行によって、消費者は当たり前のようにインターネット上で日常生活や日々考えていることを語り始めた。企業にとって消費者の個別ニーズや文脈を検出することは非常に重要であり、マスニーズだけを追い続ける企業は生き残りが難しくなっている。本稿では、消費者の多様な生の声を低コストかつリアルタイムに取得する試みとして、日本語圏ブログのリアルタイム分類とトピックの多重性に関して報告する。筆者らの実験により、ブログ記事の約半数が複数のトピックを持つことが示された。今後、より細かいブログ圏の生態と動的変化の検出が期待される。

キーワード：ブログ、ロングテール、文書分類、ナイーブベイズ、多重クラス分類

The recent boom of blogging is a phenomenon of consumer generated data in a massive scale. Simultaneously, it is becoming evidential that an enterprise whose strategic focus is only on needs at a mass-level would not survive. This paper attempts to detect individual needs at a low cost in real-time, in a form of the prompt multi-topic classification of Japanese blog entries. The results show more than half of blog entries have plural contexts. It is expected that the present method will reveal the ecology of the blogosphere and will identify its dynamics.

Keywords : blog, The Long Tail, document classification, naive bayes, multi-class categorization

1．はじめに

本稿では、インターネット上に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1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。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（自分のブログを開設しているインターネットユーザ）は延べ約335万人（複数のブログサービスへの掛け持ちを考慮すると、純ブログ利用者数は約163万人）。アクティブブログ利用者（ブログ利用者のうち、少なくとも月に1度はブログを更新しているユーザ）は約95万人。
- 2007年3月末にブログ利用者数は述べ約782万人、アクティブブログ利用者数は約296万人に達すると予測。

となっている（図1）。また、筆者らの実測によると2005年10月の時点で、朝刊200日分の文字数（5,000万文字）を越える日本語のブログが毎日書かれている。

多数の生活者の生の声がブログの形で大量に蓄積されつつあり、既に人口の無視できない部分が自ら考えていることや感じたことをインターネット上に発信する状況を、梅田は「総表現社会」[3]と称した。

「総表現社会」の1つの特異性は、アンダーソンが名付けたロングテール分布[4]の長大な尻尾部分についての情報を、インターネッ

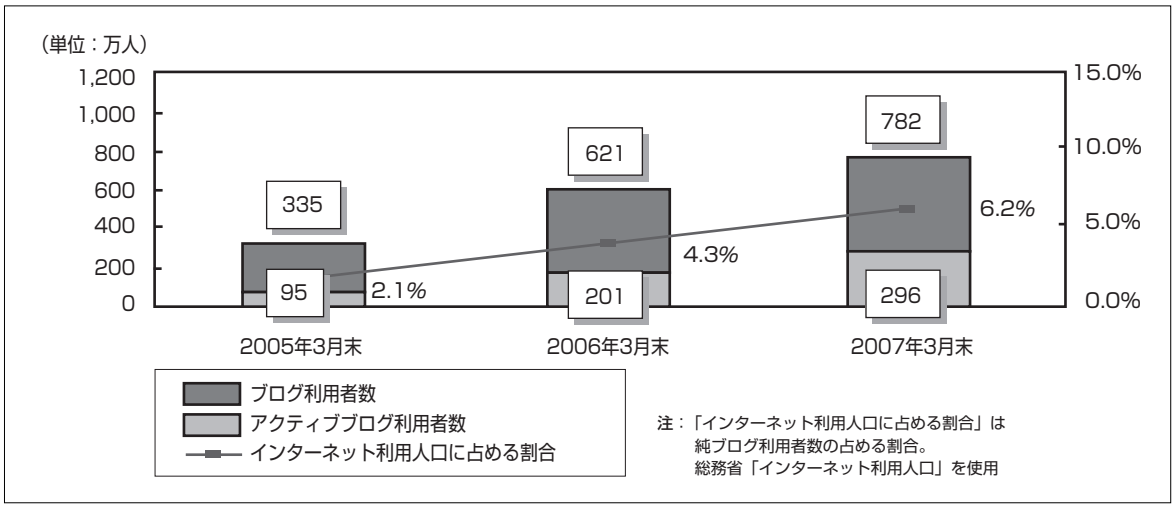


図1　ブログ利用者数（ブログ利用者・アクティブブログ利用者数）

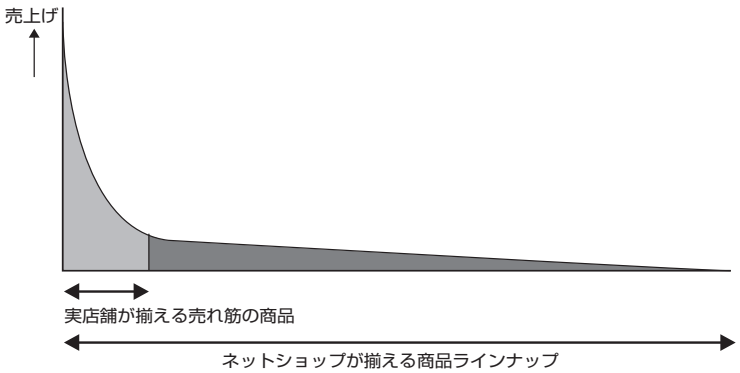
出所：総務省「ブログ・SNSの現状分析および将来予測」平成17年5月

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。ロングテール*とは、頻度分布の尻尾が長大である統計的性質を言い表したもので、頻度が下位の部分も集積すれば相当量になることを示唆している。その部分に着目した成功事例もいくつか報告されており、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ロングテールの意義がある。一般的には、ニーズの個別性をいかに低コストで検知し、それに対応するかが課題となる。

個別のニーズを検知する課題は容易ではない。この課題に対する1つのアプローチとして、筆者らはブログエントリの多重トピック自動分類を取り上げる。例えば、ブログエントリ本文が「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、一日中ほとんどけんかなしでよく遊べるなぁ。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、「出産・育児」と「料理・レシピ」といった2つのトピックに分類し得る。この例のように、1つのブログエントリが2つ以上のトピックを持つこと、言い換えれば生活者が複数のトピックを持つことは一般的なことである。このことを本稿では多重トピックと呼ぶ。仮に分類数を100とする分類体系において、ある生活者のある日のトピ

***ロングテール**

ロングテール（The Long Tail）は、オンライン書籍販売のAmazon.comやオンラインDVDレンタルのNETFLIXなど、従来とは違ったネット上の成功ビジネスモデルを説明するために、米Wired Magazine誌の編集長であるクリス・アンダーソン（Chirs Anderson）によって2004年頃から提唱された概念である。小売業における従来の概念は、売上品目上位の20%が売上高の80%を占めるという「パレートの法則」であった。ところが近年、情報技術やネットワークインフラの普及およびセルフサービス化により個品の販売コストが減少し、売れ筋商品とそうではない商品の販売コストの差は少なくなってきた。販売コストが激減した業態では、売上品目の中位から下位を集積することで従来以上の売上高比率を占め、新たな成功モデルとなってきた。この領域やこの領域重視の戦略のことを説明する場合にロングテールという言葉が用いられる。



ックを3つ特定することは、約100万ある可能的な多重トピックの順列のうち1つに絞り込んだことを意味し、かなり個別的に生活者の声を捉えたことになる。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得る。

ロングテールの議論が示唆するもう1つの点は、時間軸上での動的変化を伴うということである。どんな新しい話題も、最初は頻度分布の尻尾部分で産声をあげ、その進化過程はさまざまである。ある話題は、尻尾部分に誕生したと思ったらすぐに消滅してしまうものもあるし、ずっと尻尾部分に存在し、時間が経っても消えることなく安定に推移するものもあれば（ニッチな話題）、頻度分布の中を急速に駆け上がっていくものもあるだろう（いわゆる口コミヒット）。このような動的変化を検出することがもう1つの課題となる。口コミ的な広がりを持つ話題に関心のある事業家は、ニッチな話題がマスな話題に転じるところを検出する必要がある。そのためには、特定時点のスナップショット的な分析よりも、話題の時間的な発展過程の分析が重要になる。この時間的分析の際に、リアルタイムなブログ分類が1つの手段を提供する可能性がある。

ブログの分類に関連した最近の動向として、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、「自己申告」の情報を参照し、さまざま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の中立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) より意味の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) 分類ごと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。本稿では、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1つの速効的取り組みについて論じる。

本研究での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節「ブログエントリの自動分類手法」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節ではある期間の日本語圏ブログの分類実験の結果と考察を述べる。4節では本研究で得られた結果をまとめ、今後の研究の拡張方向性を述べる。

2．ブログエントリの自動分類手法

(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ており、体系見直しの間隔も短く、各ポータルサイトでさまざまな構成となっている。

本研究では、インターネット上に現れた多数の生の声を対象にしていることから、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。

システムが自動分類を行うにあたり、あらかじめ各カテゴリの基準となる訓練データをシステムに与えて学習させる必要がある。この初期訓練データもより生活者の視点に立ったものを与える必要がある。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)ブログ自動分類器の構成

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、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。言い換えると、個々のカテゴリに属するかどうかをベイズ法で判定し、この判定を分類体系中のすべてのカテゴリに対して適用した。

この手法を採用した第1の理由は、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ことである。巨大掲示板の記事群に

第1階層	第2階層	第1階層	第2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		

表1　本研究における分類体系

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2の理由は、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ことである。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 $\{w_1, ..., w_n\}$ 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、この文章がカテゴリ c_i に含まれる事後確率を、

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

とする。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。なお、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]。

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2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7,385個の掲示板を選び、第2階層の91カテゴリに結びつけた。この時、1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1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。

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に対する処理速度からの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数台用いて処理することで、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1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。この処理時間のほとんどは、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、$\vec{c}=\left(c_1,...,c_j\right)$とすると、式(1)は、

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

となり、式(2)は、

$$P\left(\vec{c}\right)=\left(\frac{\prod_{k=1}^{n}P(w_k\,|\,c_1)}{\prod_{k=1}^{n}P(w_k\,|\,c_1)+\prod_{k=1}^{n}(1-P(w_k\,|\,c_1))},...,\frac{\prod_{k=1}^{n}P(w_k\,|\,c_i)}{\prod_{k=1}^{n}P(w_k\,|\,c_i)+\prod_{k=1}^{n}(1-P(w_k\,|\,c_i))}\right) \cdots(4)$$

と表せる。このことにより、2値問題自動分類を91回繰り返す場合と、1度に91の2値問題自動分類を実行する場合とで、確率情報をディスク装置からCPUへ転送するデータ量は変わらないが、転送回数は91分の1ですむ。このデータ転送回数の最小化が、速度性能に大きく寄与する。

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。

多くのブログ作者は新規にブログエントリを書いた時に外部のサーバに自身のブログが更新されたことを知らせる情報を送る。この送信情報を受け付けるサイトは一般に更新情報pingサーバと呼ばれ、第三者が更新情報のリストを参照することが可能である。国内では最も多くのブログ作者がping.bloggers.jp

に更新情報を送っている。

本研究においても、ping.bloggers.jp から更新情報 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。

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、ブログ作者が書いたエントリ本文だけでなく、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いる。また、RSS フィードには先頭 100 文字程度の本文記述が含まれている。

余分な記述を含んだ html ドキュメントや本文の一部分を自動分類器への入力とすることは分類精度上好ましくない。また、生活者の主張の大きさは文字数とも関連があると考えられるため、html ドキュメントから切り出した全文の取得が必要となる。

現在のブログは書式がさまざまであるため、エントリの本文を正確に切り出すことは単純作業ではない。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。

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、という観点からいうと、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、さまざま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さまざま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経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。

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

（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。この数は対象期間中に書かれた日本語ブログの半数をカバーしていると考えられる。

なお訓練データには、Yahoo!掲示板から、7,385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。

（2）実験結果

①分類分布

対象期間中のブログエントリの第1階層分類ごとの頻度を図2に示す。ブログエントリの総文字数に対する各カテゴリの総文字数の割合を図3に示す。図3で、すべてのカテゴリの割合を足し合わせると137%になってしまうのは、あるエントリが複数のカテゴリに分類された場合にも文字数をそれぞれのカテゴリで足し合わせているためである。これはあるブログエントリが二重、三重のトピックを持っていたとして、それぞれの分類の文字数を単純に全体の1／3を割り当てることは新たな理論付加が必要なためである。

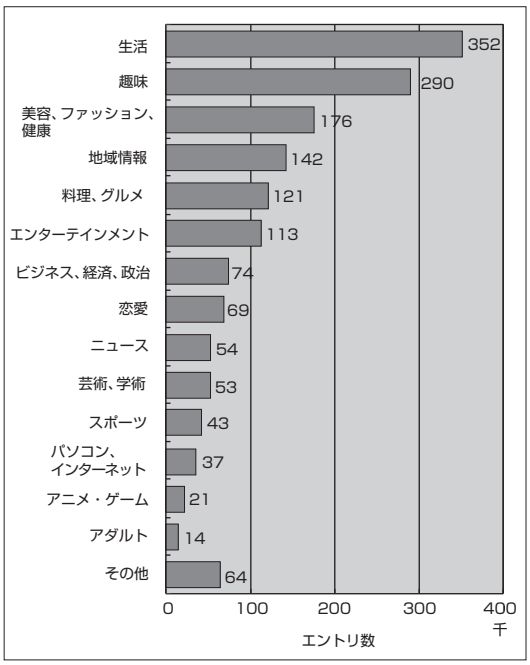


図2　1週間分のブログエントリの自動分類結果
（2005年10月1～7日）

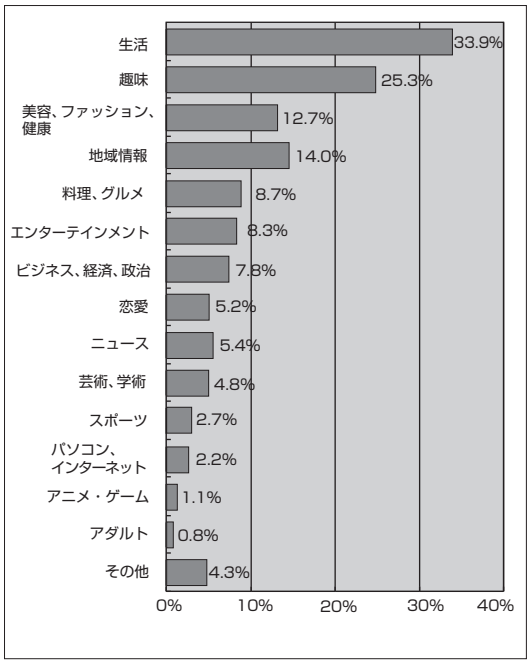


図3　ブログエントリの総文字数に対する各カテゴリの総文字数の割合
（2005年10月1～7日）

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。

1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使いも得なかったが、復路ではフライトアテンダントが温かい気使いをしてくれ、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捕えられた。

③自動分類の速度

本研究の実験環境においては、2.(2)で述べた自動分類機構は、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、1エントリーあたりの平均文字数は約490文字であった。

④トピック多重度

1つのエントリが何重のトピックを持つかを表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。

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。

⑥トピックの時間発展

1節で述べたように、ロングテールのダイナミクスについての情報として、トピックの時間発展がある。まず、日本語ブログ圏の半分以上の規模にあたる1日ごとのブログエントリ数と総文字数の推移（2005年10月1日から1週間）を図4に示す。文字数で規模を計測するとすれば、期間中すべての日において

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

表2　上位10のトピック対

5,000万文字以上生成されていたことは、全国紙朝刊が約25万文字程度（1行11文字×78行×15段×40ページ÷2（広告や写真のスペース）＝ 257,400文字）であるから、1日あたりの日本語ブログ圏の規模が朝刊200日分を超えていることになる。なお、10月3日にエントリ数・文字数に落ち込みがあるのは、参照していたpingサーバが数時間停止していたためである。

図5に、91ある第2階層カテゴリのうち例として「出産・育児」、「政治・経済」、「旅行」の3カテゴリの日ごとの文字数推移を示す。本稿が採用した分類体系第2階層の粒度では、各カテゴリで日ごとに文字数が大きく変化するものではないことがわかる。1番多い出産育児が1日1,600万文字前後であるから、新聞64日分、政治経済では新聞4日分前後が毎日生成されていたことがわかる。一方、全体の総文字数に対する各カテゴリごとの文字数の割合を算出し、2005年10月1日のその値を100として指数化したものを図6に示す。この方法により、話題の変化がより明確に可視化できる。1週間という期間で結論めいたことを論ずるのは躊躇するべきことだが、図6で示されていることは、我々の生活感覚に一致する（週末に出産育児が低く、日曜日に政治経済が大きく落ち込み、旅行が日曜日に大きくなっている）。表2と図5で得られた情報をさらに総合すると、育児中の女性がブログ圏で大きなプレゼンスを持っていることが類推できる。

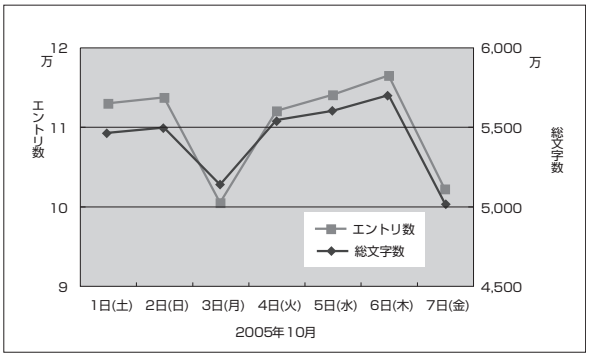


図4　実験収集したブログエントリの数と総文字数の推移

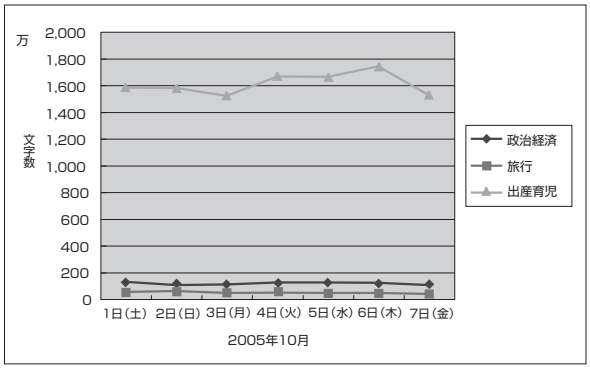


図5　各カテゴリに分類されたブログエントリの総文字数の推移（抜粋）

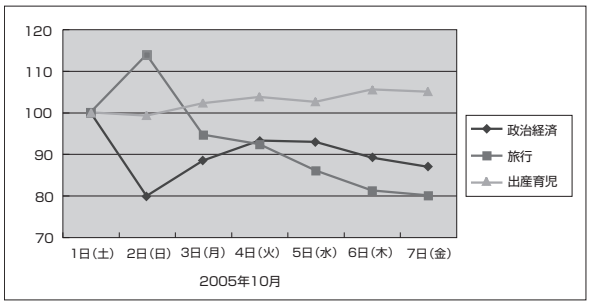


図6　収集されたブログエントリの総文字数に対する各カテゴリに分類されたブログエントリの総文字数の割合の推移
（10月1日を100として指数化）

(3) 実験のまとめ

1節「はじめに」おいて、生活者の視点によるブログのリアルタイム自動分類に関しいく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) 分類ごと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1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2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。

また、自動分類結果の時間発展に注目することで、トピックの進化、変動過程を定量的に得ることが可能であることを示した。

今後の技術的な課題として、SVMなど他の手法を用いた自動分類の精度改善と、利用者からのフィードバックに基づいた評価を分類結果に反映させるメカニズムの研究が上げられる。

4. 最後に

ブログ記事の分析によって消費者意識の側面を捉える研究は進行段階ではあるが、本実験により、日本語圏ブログのリアルタイムな自動分類が低コストで実施可能であること、多くのブログエントリが多トピックを持っており背景にある潜在的な文脈の検知が可能であること、継続的なブログ分析により話題の変化を捉えることが可能であることが示された。

ブログの自動分類が低コストで可能ということは、今後ますます日本語圏ブログの生成速度が上がったとしても、分類結果自体やその分析結果を多くの人や企業で利用することが可能であることを意味している。そして第三者の中立的な自動分類は、「自己申告」でのタグ付けやFolksonomyでは得られない潜在的な情報を検出可能であることを本実験が示している。また、本実験の1つの応用として、ブログエントリの総文字数や各分類での総文字数を1つの定量化法とし長期継続的に測定することで、株価における「日経平均」のような指標を手にすることが可能となる。

ブログの普及に伴う総表現社会の規模と生態は、これまで俯瞰的に把握する手段がほとんどなかったが、本研究はブログ圏に内在する生活者の情報を分析するための1つの手段を与えたことになる。

今後の研究の方向性として、継続して日本語圏ブログの自動分類を実施し、安定的に公表することで外部からも利用できるようにすることを中心に置き、長期計測による利用シーンの開発や特定用途向けの解像度の高い自動分類手法の開発を行い、ビジネス現場に対して有益な情報を提供する活動に取り組んでいきたい。

●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2005年5月17日付報道資料「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

［3］ 新潮社『フォーサイト』2005年6月号P.8～10「ウェブ社会［本当の大変化］はこれから始まる」梅田望夫著

［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］『第19回人工知能学会全国大会』2005年6月「Community Webプラットフォーム」大向、松尾、松村、武田著

［9］ ping.bloggers.jp
http://ping.bloggers.jp/

［10］『情報処理学会研究会報告（自然言語処理）』2002年7月P.105～110「コメントを用いた映画の自動分類」阿部、田中、中川著NL-150-16

[11] Yahoo!掲示板
http://messages.yahoo.co.jp/
[12]『情報処理学会誌 Vol.42』2001年1月 P.32～37「テキスト分類－学習理論の「見本市」－」永田、平著
[13]『情報処理学会研究会報告（自然言語処理）』2004年9月P.53～60「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」加沢、泉谷、平、前田著NL-163-8
[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]『形態素解析システム茶筌』松本他著
http://chasen.naist.jp/hiki/ChaSen/
[17]『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』2005年7月「なんでもRSS!－HTML文章からのRSS Feed自動生成」南野、奥村著SIG-SWO-A501-03
[18] ブログファン
http://blogfan.org/
[19] ニンプ沖縄に行く、著者不明
http://taratta.at.webry.info/200509/article_9.html
[20]『情報処理学会研究会報告（自然言語処理）』2005年11月P.21～26「日本語圏ブログの自動分類」平野、古林、高橋著2005-NL-170
[21]『日本語圏ブログをもとにした話題指標の開発』古林、平野、高橋著（準備中、2005）