ブログ記事の収集と予備分析
～大規模分析に向けて～

石川雅弘
池田　潔
加藤淳一

概要

この10年ほどでインターネットは一般人にとっても身近なメディアとなった。広告費という基準ではすでにラジオ・雑誌・新聞を抜き，テレビに次ぐ巨大なメディアに成長している。そこでは，一方的に情報を受信するだけではなく，誰もが発信者となることができ，双方向のコミュニケーションが可能である。また発信された情報の多くがネット上に蓄積され，時を経てからでもアクセスできる。ブログはそのような性質を持った代表的なネットメディアであり，既に多くの人々により大量のブログ記事が執筆・蓄積されており，今後も増加し続けるであろう。蓄積された大量のテキストの中には，様々な事柄に関する様々な人々のその時々の生の声が含まれていると考えられ，例えばマーケティングなどの種々の応用分野にとって有益な情報源となり得る。我々はこれまでにブログ分析のためのクラスタリングおよび可視化手法，またマーケティング分野での利用方法などについて研究を進め，その成果を発表してきた。本稿では，それら各論では触れることのできなかったブログ記事の収集方法とその基礎的分析結果について述べ，今後様々な分野でブログデータの利用を進める際の助けとしたい。

キーワード：インターネットメディア，ブログ，テキストマイニング，クラスタリング

1．はじめに

インターネットが，特別なものではなく一般の人々に利用される日常的メディアとなって久しい。2011年２月の電通による発表によると，インターネット広告費は2004年にはラジオ，2006年には雑誌，そして2009年には新聞広告費を越えた(1)。少なくともこの尺度では，インターネットはテレビに次ぐ巨大なメディアとなったと言える。また，2011年６月に発表された博報堂ＤＹメディアパートナーズによる調査結果によれば，年齢層で若干傾向が異なるものの，人々のマスメディア四種（テレビ・新聞・雑誌・ラジオ）との接触時間が減少傾向にある一方で，パソコンや携帯機器からの

インターネット接触時間は増加傾向にあり，2009年以降は新聞・雑誌・ラジオを合わせたものを越えている[2]。この尺度からも，インターネットはテレビに次ぐ巨大メディアに成長したことが裏付けられ，今後既存マスメディアのあり方にも影響を与えながら成長を続けると考えられる。

我々は，インターネット上のメディアの中でもブログに注目し，それを有効な情報源として活用するための研究を進めてきた。技術的に新規な点や主たる応用分野の一つとして想定するマーケティング分野への貢献については，それぞれ別稿で報告してきた[7][8][9][10][11]。しかし，ブログデータの収集方法とその分析手法について，特に非情報系であるマーケティング分野の研究者には自明とは言い難い点が多く追試や類似研究の実施が困難であるとの指摘を受けてきた。そこで本稿では，それらでは触れることのなかったブログデータの収集方法と基礎的な分析結果について述べ，今後様々な分野でブログデータの利用を進める際の助けとしたい。

2．インターネットとブログ

メディアとしてのインターネットには，既存マスメディアとは大きく異なる点がある。第一に，既存マスメディアが基本的に一方通行の放送・出版であるのに対し，インターネットは双方向の通信メディアである点。第二に，情報発信が少数の職業的発信者に独占されておらず，誰でもが発信者となれる点である。発信者・受信者をノードとするネットワーク構造の観点から言えば，マスメディアが特殊なノードである発信者“1”と受信者であるその他のノード“多”の1対多ネットワークであるのに対し，ネットメディアは対等なノードによる多対多ネットワークであると言える。

以上の二点の他にも，ネットメディア上の情報の多くは発信後も長くネット上に保存され，時間を経た後でも容易にアクセスできるという特長がある。中でもブログは，個人が継続的に情報発信・蓄積する場合の有力な選択肢であり，自由に文章や画像・音声などを発信することができ，またトラックバックやコメント機能により発信者と読者，あるいは読者同士がコミュニケーションを取ることもできる。そのため，単に日々の出来事や感想・考えなどを日記的に記すだけではなく，興味を共有するもの同士が情報や意見を交換するなど，コミュニティ形成の場ともなっている。ここ数年 Twitter などの比較的揮発性が高い短文メディアが台頭してきたが，そこでも重要で他者との共有が必要な情報はブログやニュースサイトなどの保存性の高い記事へのリンクとして示される事が多く，まとまった情報や意見表明におけるブログの重要性は減じていない。また，ネット上では文章以外にも画像・音声・動画などのマルチメディアの利用が進んでいるが，意見の表現メディアとして最も重要なのは依然として自然言語による文章記述であり，今後もその重要性が大きく損なわれる事はないであろう。さらに，ブログ記事には，公開日時と著者（ブロガー）情報が付随しており，ブロガー個人やその集合を時間軸に沿って分析することも可能である。ブログデータは，人々の関心や感想・意見，そしてその変化を知るための重要な情報源となり得るのである。

しかし，新聞記事のような一部の職業記者により執筆されるものとは異なり，ブログなどの文章はバックグラウンドの異なる不特定多数の一般人により，細則や校閲などがない自由・勝手な環境下で執筆される。そのため，誤字脱字も多く，また多様な同義語，短縮表現，隠語など，表記のゆ

れも大きいと考えられ，新聞記事など素性の良い文章を対象として開発されてきた従来の自然言語処理技術では対処できない問題も潜んでいるであろう。データ量でも，ブログは新聞を遥かに凌駕している。2005年に産経デジタル社が発行した産経新聞ｅテキストの記事数は約８万件である[4]。一方，goo ブログだけでも，一日に２万件以上の記事が投稿されている。また，総務省の調査によると，2008年時点の国内のアクティブブロガー数は約300万と推計されている[3]。これらを元に考えると，低く見積っても，記事数ベースで既に新聞一紙の数百年分に相当するブログ記事がネット上に蓄積されていると考えられる。パーソナルコンピュータの処理速度，メモリ容量，二次記憶容量などが長足の進歩を遂げたとは言え，これだけ大規模なデータを効率的に保存・管理・処理するには，データの保存方式や処理アルゴリズムについても慎重に検討する必要があろう。

インターネット上にはブログ記事以外にも様々なテキストデータが蓄積され続けており，そこには従来の定型アンケートでは得られない，人々の生の声が溢れていると考えられる。それらを有効な情報源として活用するには，上述の問題も考慮しながら収集と分析を行なう必要がある。

３．ブログデータの収集方法

本節ではブログ記事テキストの収集方法について述べるが，その中にはブログサイト依存の部分がある。これはブログサイトによりブログページの構造が異なり，後述する RSS や記事一覧，あるいは記事テキストを抜き出すために必要な手がかりが異なるからである。また，それらの手がかりを得るには，人の手でブログ構造を調査する他ない。そのため，複数のブログサイトを対象としてデータ収集を行なうには，それぞれのブログサイトを調査し各サイト用のプログラムを開発する必要がある。さらに，対象サイトの構造が変更された場合にはそれに応じてプログラムも修正する必要がある。

（１）対象としたブログサイト

そのような手間を軽減するため，我々は研究の目的上十分なブロガー数とブログ記事数が確保できる一つのブログサイトのみを対象とした。対象としたのは goo ブログ[5]である。goo ブログは2004年３月に開始されたブログサービスサイトであり，他の多くのブログサイト同様誰でも無料で開設でき，一定容量までは無料で利用できる。2003～2005年は，現在も続く主要な日本語ブログサイトが登場した日本におけるブログ普及期であり，goo ブログを対象とすることで初期からのブロガーも補足できると期待できる。また，ブロガー層に特に際だった偏りもなく，サンプルとして相応しいと考えた。

以下では goo ブログを対象としたブログ記事データの収集方法を述べるが，他のブログサイトを対象とした場合でも処理の概要は同じであり一般性を失なわない点に注意されたい。

（２）収集方針

ニュースサイトやブログサイトのように時間とともに更新され新たな記事が公開されるサイトで

は，新着記事を知らせるために RSS を提共するのが一般的である。RSS とは，公開日時，記事の見出し，記事の URL などを機械可読な XML 形式で公開するものである。機械可読なため，プログラムにより定期的に自動受信し，新着記事の有無を自動的にチェックすることができる。そのような目的のソフトウェアは RSS リーダーと呼ばれ，ニュースやブログ記事を効率的に閲覧するためのソフトとして普及している。読者にとっては新着記事の無い時に無駄に閲覧行動を取る必要が無くなり，サイト運営側にとっては無駄なトラフィックを抑制しつつ新着記事の閲覧を促すことができる効果的な仕組みである。

ブログデータの収集にもこの仕組みを利用できる。すなわち，新着記事を知らせる RSS を自動チェックし，新たな記事があればそれを収集すれば良い。しかしこの場合，収集できるのは収集開始以降に公開された記事のみである。過去から現在までの時間軸に沿った分析をするためには，すでに蓄積されている過去の記事が必要であり，別のアプローチが必要となる。そこで我々は，次のような手順でブログ記事を収集することとした。

1．新着記事 RSS によるブログ ID の収集
2．各ブログの記事 URL 一覧の収集
3．各ブログの全記事の収集

以下では，我々がターゲットとした goo ブログを例にとり，各手順の詳細を述べる。

（3）RSS によるブログ ID の収集

上述のとおり，RSS をチェックすることで新たなブログ記事の URL を得ることができる。2011年11月時点で，goo ブログでは次の URL で新着記事 RSS を公開している。

```
http://blog.goo.ne.jp/portalrss/recententry/
```

RSS の中身である XML データをダウンロードするには，HTTP プロトコルで上記 URL の取得要求（GET）を行なえば良い。現在多くのプログラミング言語にはインターネットで一般的なプロトコルに従って通信するための簡便な手段が用意されているため，比較的簡単なプログラムで実現することができる。我々のシステムの実装には Python 言語とその標準モジュールである urllib2 を用いた。

ダウンロードした XML データでは，各エレメントはタグでマークアップされている。どのエレメントがどのようなタグでマークアップされているかもサイト依存であり，人手で確認する必要がある。goo ブログの場合，例えば日付は

```
<dc:date>2011-11-11T18:12:21+09:00</dc:date>
```

となっており，また新着記事へのリンクは

```
<rdf:li rdf:resource="http://blog.goo.ne.jp/[blog id]/e/[entry id]"/>
```

のようになっている。ここで [blog id] にはブロガーを識別するブログIDが，[entry id] には記事を識別するエントリーIDが入っている。従ってここから [blog id] を抜き出す事で，新たに記事を公開したブログのIDが得られる。

gooブログでは新着RSSは数秒ごとに更新されるため，一定期間，数秒おきにRSSを監視することで，現在もブログを更新しているアクティブなブログのリストが得られる。

（4）各ブログの全ての記事のURLの収集

ブログには最新の記事だけではなく過去の記事も蓄積されており，過去記事を一覧できるページが用意されているのが一般的である。gooブログにもそのようなページがあり，個別記事のページやトップページ（http://blog.goo.ne.jp/[blog id]/）からもリンクが用意されている。しかしgooブログでは記事一覧ページのURLは

```
http://blog.goo.ne.jp/[blog id]/arcv/
```

であり，ブログIDが入手できれば容易にアクセスできる。ただし，蓄積記事数が多い場合には，ネット検索での結果表示と同様に複数ページに分割されて表示されるため，リンクを辿って複数のページをチェックする必要がある。

記事一覧ページのHTML中では，各記事へのリンクは エレメントの <ul> エレメント（番号無し箇条書き）内の （箇条書の項目）タグで

```
<li><span class="mod-arcv-tit"><a href="http://blog.goo.ne.jp/[blog
id]/e/[entry id]">[title]</a></span><br />
```

のように記述されているため，<a> タグの href 属性を取り出せば良い。ここでコンテンツ [title] はその記事のタイトルである。また，記事一覧を含むエレメント以降の <a> エレメントでコンテンツが“次へ”のものがある場合は次ページが存在することを意味し，次ページへのリンクはその <a> タグの href 属性に設定されている。したがって，“次へ”を辿りながら記事URLを取り出すことで，全ての記事のURLを得ることができる。

（5）ブログの全記事テキストの収集

ここまでで，現在も更新が続くアクティブなブログの全記事のURLを手に入れることができた。最後に必要なのは，各記事のHTMLデータをダウンロードし，そこから記事テキストのみを取り出すことである。

gooブログでは，HTMLデータの エレメントのコンテンツとして記事テキストが記述されているため，不要なHTMLタグなどを削除しつつ抜き出せば良い。記事のタイトルと公開日時については，rdf:Descriptionタグの dc:title 属性と dc:date 属性にそれぞれ記述されている。

図1にブログ記事収集の手順を，図2に収集システムの構成イメージを示す。

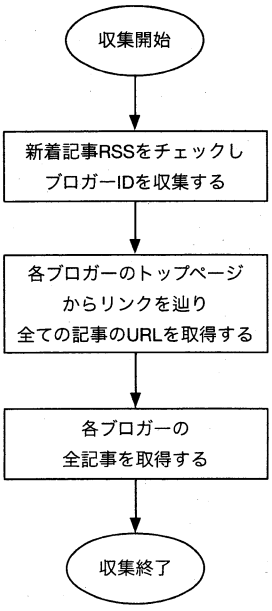

図1　ブログデータ収集の手順

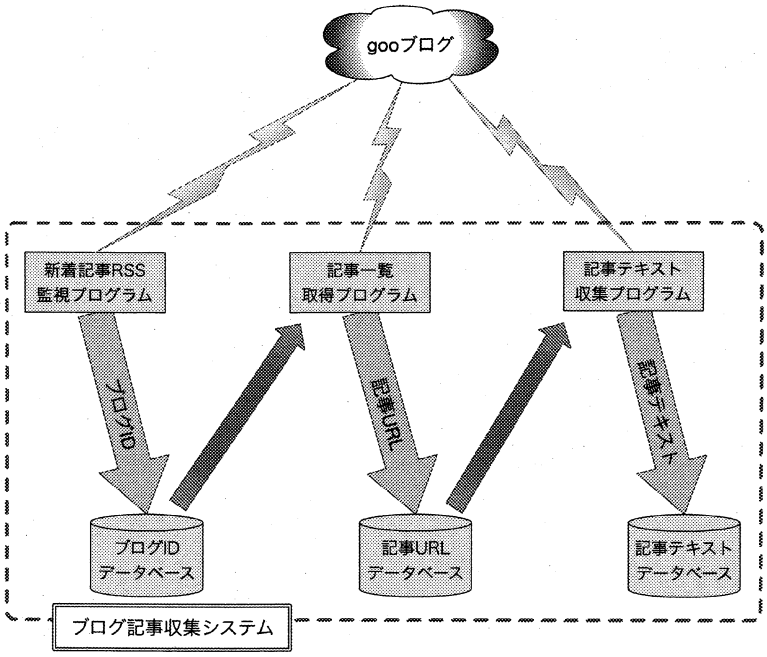

図2　システムのイメージ

（6）収集にあたっての注意

前節までにブログ記事の収集手順の概要を述べた。しかし実際に収集作業を行なう場合には，さらに細かい点での注意が必要である。

①日本語文字コードの問題

現在コンピュータ上で使用されている日本語文字コードは複数種類あり，それらを適切に扱わないと文字化けを起こしたり，同一文字列でも異なるものと判断してしまうなどのミスが発生し得る。また，HTMLではメタ情報で文字コードを指定できるが，必ずしもそれが正しいとは限らない。そのため，万全を期すには日本語を含むデータをダウンロードした場合には文字コードの判定を行ない，目的の文字コードに変換するのが望ましい。処理・保存するデータを一つの文字コードに統一することで，その後の問題を避けることができる。統一文字コードとしては，現在標準的な地位を固めているUTF-8が好ましい。ただし，他の文字コードでは存在するがUTF-8では存在しない（定義されていない）文字があるため，文字の置き換えなどそれらへの統一的対処が必要である。また，全角アルファベットを半角アルファベットに，大文字を小文字にするなど文字種の統一も検討する必要がある。

②並列化と相手サイトへの負荷

図1に示す収集手順では三つのプロセスは順番に行なう必要があるように見えるが，図2に示したように，実際にはこれらのプロセスは並行して実行することができる。また，既に複数のブログIDが得られている場合，複数のブログの記事を同時並行的に収集することができるなど，収集手順には様々な並列化可能性が潜んでいる。一般にネットアクセスには，データを転送する時間以外に相手の応答を待つ時間があり，プログラム実行中でも必ずしもCPUパワーや通信バンドの全てを使っているわけではない。そのため並列度を上げることでスループットを高める工夫が必要となる。また，近年のCPUは多コア化が進み一台のCPUでも同時に複数の処理を実行できるため，計算機資源を活用するにはより一層の並列化が必要である。

ただし，アクセス先への負荷も考慮する必要がある。あまりに高頻度なアクセスの長時間に渡る継続は，対象サイトに高い負荷をかけることになり，当該サイトのサービス提共に支障を生じさせないとも限らない。また，迷惑アクセスと判断されアクセスをブロックされることもあり得る。複数のブログサイトをターゲットとする場合には，一つのサイトへの同時アクセスは一つのみとし，並列化は複数サイトへの同時アクセスにより実現することでこの問題は軽減できるが，今回のように一つのサイトのみを対象とした場合には注意が必要である。同一サイトから短期間で大量のデータを収集する場合には，相手サイトから承諾を得たり，可能ならば別の形でデータ提共を受けるなどの配慮も必要であろう。

4．ブログ記事の収集と基礎的分析

これまでに述べたような方法で，2011年と2012年の二期に分けてgooブログを対象としたブログ記事データの収集を行なった。ここではそれぞれで収集したデータの概要と他稿では触れなかった基礎的な分析結果を述べる。

（1）第一期収集データ

2010年5月9日の24時間に渡り新着RSSを監視することで29641個のユニークなブログIDを収集した。次いでそれらのブログについて，全記事データの収集を試みた。ネット接続とは不安定なものであり，閲覧可能な記事でも何らかの理由でダウンロードが失敗することがある。その場合には何度か再試行を行なったが，最終的に失敗するものもあった。また，記事に付随する日付データがgooブログ開始以前になっているなど，不正なデータも見られた。

第一期に収集対象としたブログと記事の数を表1に示す。ダウンロードしたHTMLファイルの総量は約330GB，そこから抜き出しUTF-8に変換した記事テキストは約70 GBとなった。2005年に産経eテキストで配信された全記事の本文テキストは約320MBであり，その200倍を超えている。なお，大規模な収集作業ははじめてだったこともあり，並列化の程度も低くまたプロセスをチェックしつつ断続的に行なったため，開始から終了まで約二ヶ月を要した。

表1　第一期（2010 年）に収集したデータの概要

	対　象　数	全ての記事が収集できたもの	異常日付の無かったもの
ブロガー数	29,641	25,217	25,198
記事総数	15,964,628	13,960,651	13,461,962
平均記事数	592	554	534

（2）第一期収集データの基礎的分析

全ての記事が収集され，日付データにも異常のなかった25198ブログの記事データを対象に行なった基礎的な分析結果を示す。

①ブログ開始年月日

図3は，2004年4月1日を起点として各日にいくつのブログが開始されたかを示すものである。図から開始件数が飛び抜けて多い日が定期的に出現するのが見て取れるが，これは各年の元日であった（図中に日付を付した）。例外的に2009年9月21日は，元日を大きく超えている。2009年の9月は，土曜・日曜と敬老の日，秋分の日，そしてハッピーマンデーの適用が連なり，19日土曜日から23日水曜日まで秋の大型連休が実現した。21日はその中日であったことが関係していると考えられる。

最後の半年ほどはブログ開設者が急増しているが，この解釈には注意を要する。ブログを開設すると，その後しばらくの間は熱心に更新されるが，時間が経つにつれ更新されず放置されるブログが増えると考えられる。従って，早くに開設されたブログほど現在もアクティブである確率は低く，

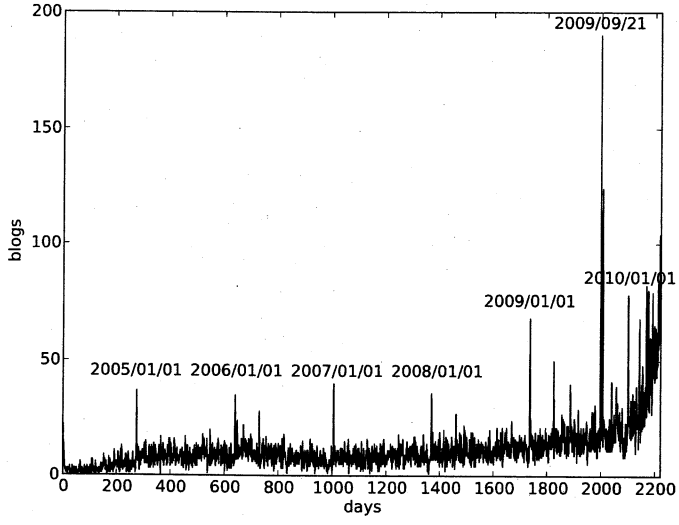

図3　日ごとのブログ開設者数

逆に最近開設されたブログほどその率は高いと考えられる。図3に見られる最新半年ほどのブログ開設数の急増はこの事情を反映したものと考えられ，単純にブログ開設者が増えたと捉えることはできない。また，半年程度継続されているブログならばその後も継続される可能性が高いと考えることもできる。

図4は2004年4月1日を起点として各日に公開されたブログ記事がいくつあったかを示すものである。ブロガーの増加に伴ない確実に増加してきたことが分かる。また，最後の急上昇はブログ開設者の急上昇と同じ理由と考えられる。

②キーワード分析

自然言語処理における文法構造や意味の解析技術は未だその信頼性は低く，表記のゆれの大きなテキストデータに適用するのはさらに難しい。そのため，ブログを含むテキストデータの分析ではbag-of-wordsと呼ばれるキーワードの集合を用いる事が多い。これは情報検索や文書クラスタリングなどでも一般的な形式で，各テキストデータを，それが含むキーワードの集合（要素の重複出現を許すため，正確にはマルチ集合すなわちbag）で表すものである[12]。

そのためにはまずテキストからキーワードを抽出する必要がある。ここでは形態素解析器MeCab (v0.97)[6]により形態素解析を行ない，名詞と判定された語をキーワードとして抽出した。ただし代名詞，非自立語，接頭辞，接尾辞，数詞，サ変接続詞は除いた．なお，MeCab用辞書としてはIPA辞書を用いた．形態素解析の結果は辞書に大きく依存するため，他の実験との比較分析を行なうためには辞書の統一が必要である。

対象とした13,461,962のブログテキストにMeCabを適用した結果，4,853,498のユニークなキーワードが抽出された。記事平均で約0.36個の新たなキーワードを使用していることになる。

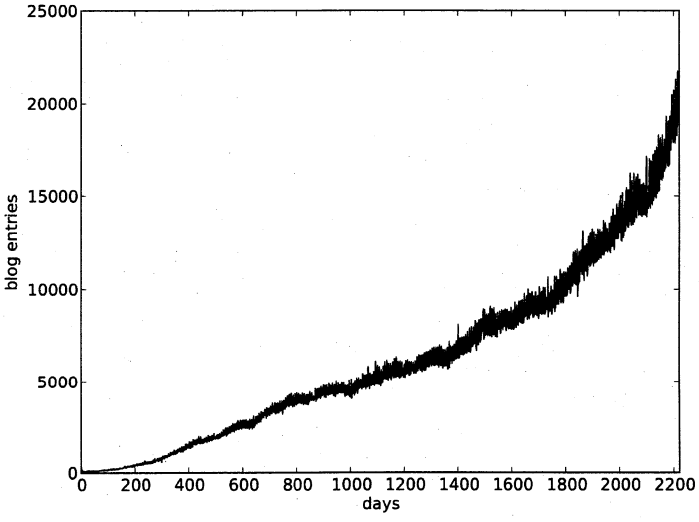

図4　日ごとの記事数

■記事数とキーワード数の関係

2005年に発行された産経ｅテキストでは，記事数は73,388，総キーワード数は108,750，記事当たり新語数は1.48である。この数値だけを見ると，ブログ記事の方が新聞記事よりも新語の登場比率は低いようであるが，単純にそうであるとも言えない。

図５に，記事数の増加に伴なうキーワードの異り数の増加の様子を示す。横軸が記事数，縦軸がそれまでに出現したキーワードの異なり数である。図から分かるように，記事数が多くなるにつれ，新キーワードの出現比率はブログの方が高くなる。これは新聞記事は統制された環境下で執筆されるため，語彙の選択が一定程度規制されているためと考えられる。それに対しブログ記事には様々な異表記や短縮表現，誤字脱字などが多いと考えられ，さらに新語の生産や導入にも制約がない。そのため，図では50万記事で打ち切ってあるが，この先もキーワードの増加傾向は変わらず，飽和する傾向は見られない。従って今後時が経ちブログ記事の集積が進むにつれて，ますます多くのキーワードが登場すると考えられる。これは後段の分析処理にとっては記憶・格納コストおよび時間コストの増大を意味し，従来マーケティング分野で用いられてきた小規模なデータを対象とした高コストな分析手法が適用不能になる可能性を示している。

■キーワードを構成する文字種

表２にキーワードを文字種で分類したときのそれぞれの比率を示す。表から分かるように，漢字のみからなるキーワードの割合が，ブログにおいては新聞記事と比べて極端に少ない。また，アルファベットのみからなるキーワードが極端に多くなっている。

漢字キーワードの平均長は，産経ｅテキストで2.281文字，ブログテキストで2.278文字とほぼ変わらない。しかしひらがなキーワードでは，産経ｅテキストで3.840，ブログテキストでは8.341，また片仮名キーワードでは6.594と8.428である。

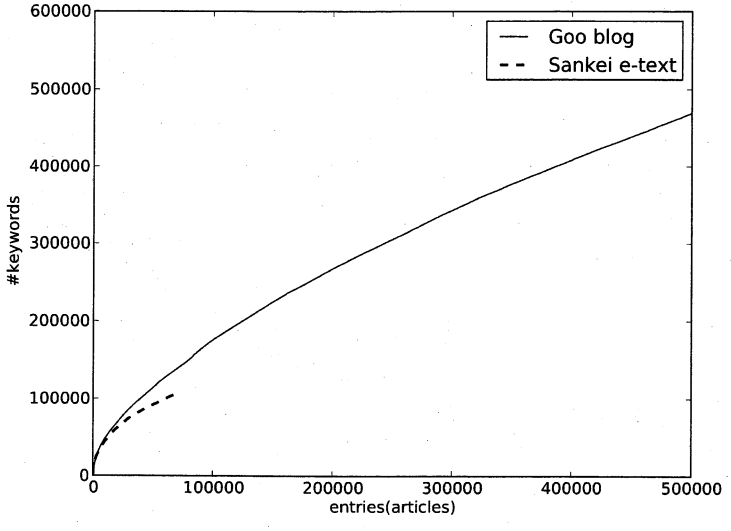

図５　記事の増加に伴なう単語異なり数の増加

表2　キーワードを構成する文字種の割合

	産経 e テキスト	Goo blog
漢字のみ	44.7%	4.1%
カタカナのみ	40.2%	50.3%
ひらがなのみ	3.6%	10.4%
ひらがなと漢字	4.6%	0.2%
アルファベットのみ	5.3%	32.4%

この差異の原因はまだはっきりとは断定できないが，データを素直に読めば，ブログでは新聞に比べひらがな，カタカナ，アルファベットが多用される傾向にあると言え，これは経験的感覚とも合致する。その他に，形態素解析器が未知語に遭遇した場合，ひらなが，カタカナ，アルファベットの一つの文字種のみから構成されている場合は名詞と判定されるケースが比較的多いためと考えられる。

（3）第二期収集データ

第一期で収集したブログのうち，2004～2007年の間に開設され，かつ2011年８月中に新たな記事を公開した7842ブログを対象に，一期収集後2011年８月31日までに公開された記事データの追加収集を試みた。そのうち全記事が収集できたのは7548ブログであった。

（4）第二期収集データのクラスター構造の時間変化

7548ブログについて，2010年９月１日～2011年８月31日に公開された記事テキストを対象にクラスター構造の時間変化の分析を試みた。分析手法は文献（9）で提案したクラスタリングとその経時変化の面グラフによる可視化であり，可視化表現を若干修正したものである（文献（9）では相対頻度の面グラフだが，ここでは頻度の波状グラフとした）。図６にその結果を示す。横軸は2010年９月１日を起点とした日数，波状の水平線はクラスタリングで得られた64のクラスターを表す。波の大きさは，その日に公開されたブログ記事中でそのクラスターに属する記事の数を表わす。したがって，波の大きさの変化はその話題（クラスター）に属するブログ記事数の時間に伴なう増減を表している。なお，64というクラスター数はパラメータとして与えたものであり，必ずしも最適なものではない。

図から，123日目（2011年１月１日）をピークに大きなクラスターが一つ徐々に現われ，また消えていく様子が確認できる。ピークの一週間ほど前にも小さなピークがあるのが分かる。さらに，一旦収束して消えた後，166日目（2011年２月14日）を中心にして小さな波がある。このクラスターは正月，クリスマス，バレンタインなどのイベントに関連した記事から構成されていると推測できる。

192日目（2011年３月11日）には，それまでほとんど存在しなかったクラスターが一つ，突然大きなクラスターとして出現し（図中破線円で示した），その後収束しつつもほぼ定常的に存在している事が分かる。これは2011年３月11日に発生した東日本大震災に関連する記事のクラスターであろ

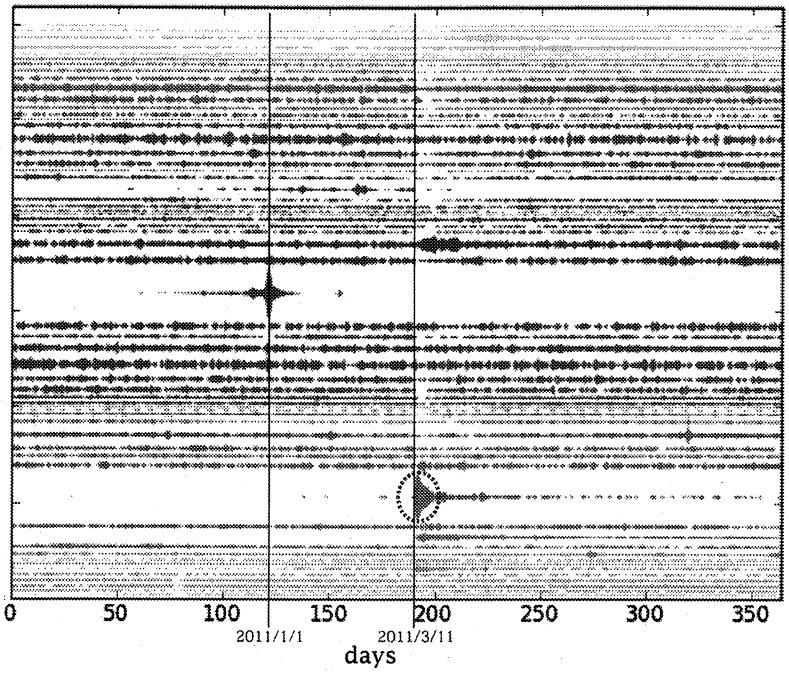

図6　64のクラスターに分割した結果の可視化例

う。

クラスターの内容を説明する情報の抽出は今後の課題であるが，このような分析は記事クラスターと実社会の出来事を結びつけ，世の中の人々が様々な出来事にどのような反応を示しているかを知るための手がかりになる。また，例えば特定の商品群に興味をもったブロガーグループを抽出しマーケット分析に役立てる事も可能だろう。

ここで示したのはブログ記事のクラスター構造であるが，ブロガーのクラスター構造とその経時変化を捉えることで，分析者の目的により適したブロガーグループの抽出が可能と考えられる。

5．まとめ

本稿では，ブログデータの収集と基礎的分析結果について，特に情報技術分野やマーケティング分野の専門的論文では触れることの少ない面に絞って解説した。データの収集については，goo ブログを例にとり，具体的な手順を示した。また新聞記事データの分析結果と比較することでブログ記事の特徴を示した。さらにブログ記事のクラスター構造の経時変化の可視化例を示し，ブログを分析することで世の中の出来事とそれに対する人々の反応などを分析できる可能性がある事を示した。

我々は現在，各クラスターの内容を説明する情報の抽出技術と，ブロガーのクラスター構造とその経時変化の可視化手法の開発に取り組んでいる。また，それらをマーケット分析に活用する手法についても考案したい。今後のブログデータの収集は，短期間に集中的に行なうのではなく，既に全記事データを収集したブログの新着記事を中心に継続的に行なう予定であり，一部すでに開始し

ている。また，英文ブログの収集も行ない，日本語と英語の違い，あるいは同一テーマに対する日本人と英米人の反応の違いなどの分析を行ないたい。

既に収集データは十分巨大なものとなり，例えばフィッシャーの正確確率検定などの適用は計算量的に非現実的となった。データは今後増加する一方であり，今後開発する分析手法はデータ量の増加に対応できるような効率的なものとしたい。

謝辞

本研究の一部は，平成22年度および23年度のつくば国際大学共同研究の助成によるものです。ここに記して謝意を表します。

（いしかわ・まさひろ　　メディア社会学科）
（いけだ・きよし　　保健栄養学科）
（かとう・じゅんいち　　メディア社会学科）

参考文献

（1）電通，「2010 年日本の広告費」，http://www.dentsu.co.jp/news/release/2011/pdf/2011019-0223.pdf，2011/02/23.

（2）博報堂 DY メディアパートナーズ，「メディア定点調査2011」，http://www.hakuhodody-media.co.jp/wordpress/wp-content/uploads/2011/06/HDYnews110615.pdf，2011/06/15.

（3）総務省情報通信政策研究所調査研究部博，「ブログの実態に関する調査研究」，2008.

（4）産経 e テキスト，https://webs.sankei.co.jp/sankei/about etxt.html.

（5）goo ブログ，http://blog.goo.ne.jp/.

（6）MeCab: Yet Another Part-of-Speech and Morphological Analyzer, http://mecab.sourceforge.net/.

（7）石川雅弘，クラスター構造の経時変化を可視化するための Time-Arrayed SOM の提案，情報処理学会第72回全国大会講演論文集 1，pp. 601－602，2010.

（8）石川雅弘，時系列テキストコレクションの可視化，電子情報通信学会，IEICE SIG Notes WI2-2010-1～31，pp. 125－130，2010年 3 月.

（9）Masahiro Ishikawa, *Visualizing Cluster Structures and Their Changes over Time by Two-Step Application of Self-Organizing Maps,* Proceedings of the 2011 International Workshop on Behavior Informatics at the 15th Paci.c-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2011), pp. 160–171, Shenzhen, China, May 2011.

（10）加藤淳一　石川雅弘，大量ブログ記事をデータとした市場セグメンテーションの半自動的分析手順，日本オペレーションズリサーチ学会2011年春季研究発表会，2011年 3 月.

（11）Junichi Kato, *Customers' Needs for Digital Terrestrial Television Broadcasting: An Analysis of Weblog*

Data, Proceedings of The 8th International Conference on Innovation and Management, pp. 1093–1096, 2011.

(12) Ricardo Baeza-Yates and Berthier Ribeiro-Neto, *Modern Information Retrieval,* Addison Wesley, 1999.

Collecting huge amount of blog entries and preliminary analysis of them

Masahiro Ishikawa, Kiyoshi Ikeda, Junichi Kato

In the last decade, the internet has become of a familiar communication media for ordinary people. Today, the internet is the second largest media (the first is the television) with reference to advertising expenses consumed. In the internet, not only "privileged" professionals but also ordinary people can express their opinions, emotions, and so on. Once information is published, most of them are accumulated on the internet and they can be easily accessed anytime. Blog (or Weblog) is one of internet media which has such property. Nowadays, many bloggers produce huge amount of articles and they are accumulated on the internet. In the huge amount of blog articles, we could find their thoughts, emotions, reactions to events happened in the real world, and so on. We think they can be utilized in many application fields, including market research, by careful analysis. In this article, we show how huge amount of blog articles accumulated on the internet can be collected and some results of preliminary analysis of them which suggests directions of further research.

Key words: Internet Media, Blog, Text Mining, Clustering