4S-5

# ブログスフィアからの最新の話題・流行の分析*

角田 恵美† 矢吹 太朗 佐久田 博司
青山学院大学理工学部情報テクノロジー学科‡

## 1 はじめに

近年 Web 上では、ブログを用いた個人による情報発信が盛んに行われている。ブログは、容易に書き込むことが可能であり、通常の Web サイトよりも更新頻度が増える。また、1 ページ 1 トピックという特徴も持ち、基本的に 1 つの話題に関して集中的に書く傾向がある。そのため、必然的にキーワードの出現率が高くなり、その記事の特徴を明瞭に示す。このことからブログスフィアでの話題情報を活用しようという試みが近年増加している。

山名ら [1] は、ユーザの検索語が含まれるブログ記事を取得し、肯定的か否定的かの判断を行い、ブログ記事を表示するシステムを提案している。これは、ユーザの検索語に対して、評判情報を効率的に収集でき、対象を選ぶ時の判断材料を増やすことができている。古川ら [2] は、ブログ間の語の伝播に注目し、議論の連なりやすい語を重要語として抽出する手法を提案した。また情報ポータルサイトである yahoo![3] や goo[4] などでもブログ上の情報を活用するためにブログ検索やブログランキングなど提供されている。しかしながら、これらは、入力したワードに対しての検索、重要語を抽出する目的、限られたブログ内だけの検索である。そこで、自動的にかつ語の出現頻度にのみ注目し重要語を抽出し、今後急速に話題となるキーワードのパターンを分析していく。

本稿では、まず自動的に主要なブログ・ホスティングサービスの最新のブログ記事から、キーワードを抽出する。次に、時間区切りでキーワードの出現頻度を算出し、キーワードの頻度が時間区切る方や時間の経過によってどのような傾向が見られるかを追い分析する。それにより、意図的にメディアなどに取り上げられた話題だけでなく、ブログ上に隠れている新たな話題について、より早く、リアルタイムで検出することが可能になる。

## 2 ブログ上のキーワード抽出、システム概要

Web 上にあるブログ記事から話題となっているキーワード（以下話題語とする）を抽出する。処理は、まず最新記事を収集し、データベースへの登録を行う。次に、データベース中のブログ記事からキーワードを抽出し話題語の決定を行う。

前者の処理では、特定のブログポータルサイト上から新しく更新されたブログサイトの情報を RSS の形式で取得し、RSS の記述を基に実際のブログサイトへアクセスすることで書き手や記事の情報を取得し、データベースに保存する。今回利用したブログサイトは、後者の処理では、投稿されたブログ記事を形態素解析システム Sen を利用し、出現頻度を計算しデータベースに保存する。出現頻度は、df(語の出現する文書数) を指標とし、時間の範囲を指定し計算する。

図 1 にシステムの構造を示す.

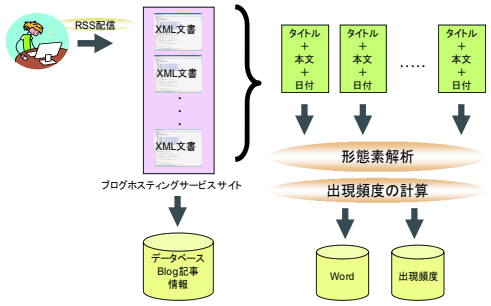


図 1 システム構造

## 3 実験概要

本実験では、話題語の出現パターンを分類することでメディアで取り上げられることのない隠れた話題をリアルタイムに取得することを目的とする。

取得した出現頻度のデータを基に、期間を限定し、急速に出現頻度が増加するデータに注目し、パターンの分類を行った。まず期間は 2007/12/1〜2007/12/7 とし、対象をブログのタイトルに絞って一週間で変化の見られる単語の推移を検出した。ただし、変化の基準をここでは単語の頻度が前 2 日の平均の 4 倍より大きい時とした。

- 本実験で扱った記事数：約 14 万件
- 単語数：約 11 万語
- 条件に当てはまった話題語：44 語

その一部を分析結果で紹介する。

---

* Discovery of the Newest Subject and Trend of BlogSphere
† Emi TSUNODA(tsunoda@idea.aoyama.ac.jp)
‡ Department of Integrated Information Technology, College of Science and Engineering, Aoyama Gakuin University

### 3.1 分析結果

本実験での分析の結果について述べる。

#### 3.1.1 話題語パターン分析

本稿では話題語パターンを 3 つに分類した [5]。検出した話題語パターンの一覧を以下に示す。

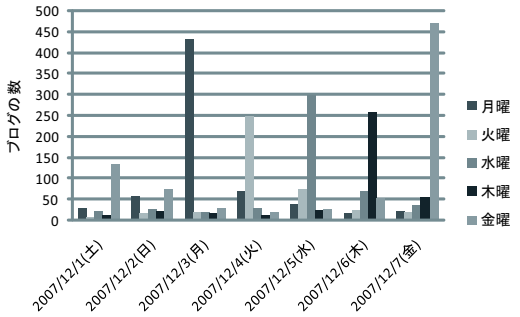


図 2　1. トリビアルな例

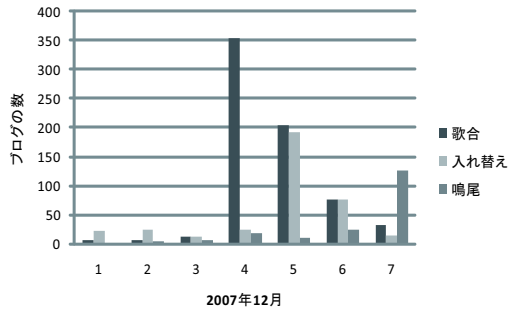


図 3　2. 大手メディアでも扱われるようなニュースを確認できた例

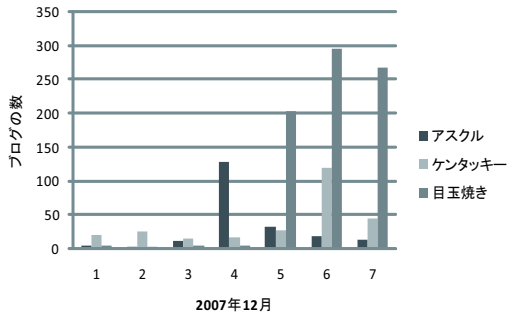


図 4　3. 大手メディアでは扱われないようなニュースを発見できた例

### 3.2 考察

本稿では、話題語パターンを 3 つに分類し、それぞれの特徴を以下に述べる。

1. 周期型

   図 2 を参照とする。月曜日にはタイトルに「月曜」と入る記事が増加する。このような周期的な出来事に関しての単語が周期性を持つパターンで現れる。

2. ニュース型

   図 3 を参照とする。4 日に急激な増加が見られる「歌合」という単語は、4 日のニュースで取り上げられた紅白出場者決定という発表によって盛り上がりを見せたと考えられる。同様に、5 日に急増した「入れ替え」という単語は、J リーグの入れ替え戦が行われたことにより、7 日に急増した「鳴尾」は、8 日に競馬の鳴尾記念が行われることにより、書き込みが増加したと考えられる。以上より、一般的によく知られたニュースやイベントにリンクする単語がニュースに並行して現れる。

3. 隠れ型

   図 4 を参照とする。4 日に急激な増加がみられる「アスクル」という単語は大手ニュースで取り上げらているわけでもない単語である。「ケンタッキー」や「目玉焼き」も同様に直接ニュースに取り上げられた内容ではない。以上より、2 のイベント型とは違い、特に大手ニュースで取り上げられているわけでもない単語が勃発的に現れる。

本実験で分類された 3 つのパターンでは、2 のイベント型の話題語によって、ニュース中でのユーザが関心を惹く話題語に関して把握することが可能になり、3 の隠れ型の話題語によって、新たに関心を惹く話題語の検出や、新たにメディアで取り上げられる可能性のある話題語を把握することが可能になると考えられる。

### 3.3 今後の課題

本文からの検出、複数単語や未知語からの検出により、隠れ型の単語をより多く検出することができると推測される。また話題語とニュースとが一致するかの判断方法を自動で行えるようにすることで、より迅速な隠れ型の話題語を検出することができると推測される。

## 4 まとめ

本稿では、ブログを用いて、話題語を抽出するシステムについて述べ、本システムを用いて行った分析結果の話題語の出現パターンににについて述べた。これにより、ブログ上に隠れている新たな話題がリアルタイムに検出することが可能になった。

## 参考文献

[1] 山名健悟, 滝沢敏裕, 湯浅将英, 大山実：Blog 記事を用いた選定支援システム, 情報処理学会, Vol. 67, No. 2U-4, pp. 3_191–3_192 (2005).

[2] 古川忠延, 松尾豊, 大向一輝, 内山幸樹, 石塚満：ブログ上での話題語伝播に注目した重要語抽出, 人工知能学会全国大会, Vol. 21, No. 2F4-4, pp. 1–4 (2007).

[3] yahoo!blog. http://blog-search.yahoo.co.jp/.

[4] gooblog. http://blog.search.goo.ne.jp/pr/index.html.

[5] 福原知宏, 村山敏泰, 中川裕志, 西田豊明：Weblog から社会の関心を探る, 人工知能学会全国大会, Vol. 20, No. 3D2-01, pp. 1–3 (2006).