

ブログコミュニティにおける話題波及の検出の試み

小阪 有平[†] 安村 禎明[†] 上原 邦昭[†]

ブログコミュニティ内で注目を集めている話題を抽出し、その波及を検出するシステムの開発を目指す。このシステムでは、ブログ記事をユーザごとに集計し、記事データからユーザをカテゴリに分類する。カテゴリごとに分類されたブログユーザ群をそのカテゴリのブログコミュニティとし、コミュニティ内のユーザが書いたブログ記事から、話題の盛り上がりを検出する。次に、話題の盛り上がりがあるコミュニティに波及することを検出する。

Detection of Topic Spread in Blog Community

KOSAKA YUHEI[†], YASUMURA YOSHIKI[†] and KUNIYUKI UEHARA[†]

We develop a system for detecting new topics and their spread in a blog community. This system classifies bloggers into categories based on the blog pages. We define the bloggers classified into one category as a community. The system detects a new burst topic from the blog pages written by bloggers in a community. Then, the system detects the spread of the new topic to the other communities.

1. はじめに

ブログは、個人が手軽に情報を発信できるメディアであり近年急速にユーザ数が増加している。ブログ記事は日々生成され、蓄積されるという性質を有するため、動的に変化する。このことは、ブログ記事がブログユーザの興味・関心をリアルタイムに反映することを意味する。ブログ記事に記述された情報を分析することは、社会的関心の把握やマーケティングデータとしての応用などに有用である¹⁾。このため、ブログ検索やブログからの評判情報の抽出が研究されている^{2),3)}。

現在使用されているブログ検索は Technorati や kizasi.jp などがあり、yahoo, google にもブログ検索専用のコンテンツが存在する。google のブログ検索は ping サーバから取り寄せたブログ記事情報に従来の Web ページの検索技術を適用したものである。Technorati, yahoo は従来の Web 検索とともに、バースト検知を利用して話題の盛り上がりを検出する技術が採り入れられ、その話題を検索ユーザに提示している。しかし、ブログ全体での盛り上がりを検出するため、共通の趣味を持つブログユーザ間でしか語られないよ

うな小さな話題は検出が困難である。マーケティングデータの観点から見ると、すでにブログ全体で話題になっている情報は、世間に対しても知れ渡っている情報である場合が多く、新たな需要を喚起するような情報ではない場合が多い。このため有用な情報は、

- 一部の人間で盛り上がりを見せている
- 全体で話題になっていない
- 今後、全体に波及する可能性がある情報

と考えられる。

kizasi.jp はコミュニティに着目して、一部のブログユーザの間での話題の抽出を行っているが、カテゴリ数は少なく、それらカテゴリ間での話題波及の追跡までは考慮されていない。

本稿では、ブログの一部で盛り上がり、その後全体に波及するような話題の抽出を試みる。このためにまずブログ記事をユーザごとに集計し、記事データからユーザをカテゴリに分類する。カテゴリごとに分類されたブログユーザ群をそのカテゴリのブログコミュニティとする。このコミュニティ内のユーザが書いたブログ記事から、意見の盛り上がりを検出する。さらに全体に波及した例を調査し、話題追跡について議論する。

technorati

[†] 神戸大学大学院工学研究科
Graduate School of Engineering, Kobe University

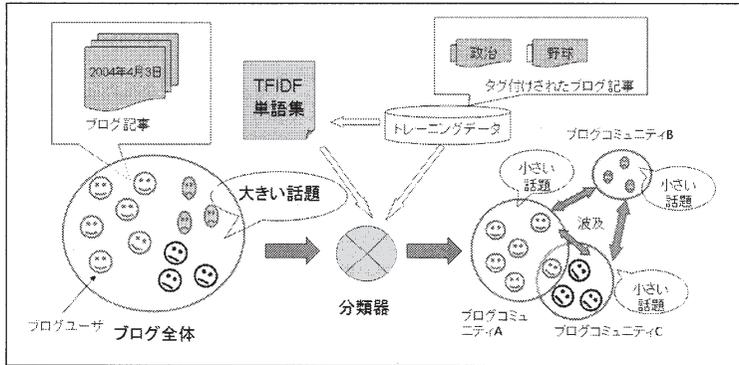


図1 システム概要

2. システム概要

本システムの概要を図1に示す。本システムの目的は、ブログ全体に波及する話題を、全体に波及する前に抽出することである。ブログ全体に波及した話題はすでに世間で知れ渡った情報である場合が多く、有用な情報でない。有用な情報は、一部の人の間で人気を集め、その後、全体に波及する話題であると考えられる。このような話題を抽出するため、ユーザをコミュニティに分け、話題の抽出を行う手法を実装する。

図1に示すように、本システムはユーザ全体をカテゴリごとに分類し、ブログコミュニティを生成する。ユーザの分類は、ユーザが過去に書いたブログ記事から判定する。記事の分類には、学習を用いる。この学習のトレーニングデータは、ブログコミュニティのカテゴリを手動で付けたブログ記事である。このトレーニングデータからTFIDFを算出し、カテゴリを代表する単語を抽出する。さらに、その単語を基に分類器を生成する。この分類器を使って実際のユーザを、カテゴリごとのブログコミュニティに分類する。

ブログコミュニティが生成できると、コミュニティごとに話題の抽出を行う。話題の抽出は、ブログの本文に出現する単語がバースト状態にあるかどうかで判断する。また、コミュニティの話題が、他の話題に波及するかの調査をする。コミュニティ C_1 が単語 W でバーストし、時系列的に後に、コミュニティ C_2 で単語 W がバーストすれば、話題は C_1 から C_2 に波及したとする。

以下では、それぞれの手法の詳細を説明する。

3. ブログコミュニティの生成

ここでは、ブログコミュニティの詳細とその生成手法について述べる。

3.1 ユーザの専門性

ブログユーザは、ブログ記事を自由意思のもとで作成しているが、その内容は、ユーザ自身の趣味に依存している。例えば、野球好きのユーザはプロ野球を題材にブログ記事を書くことが多く、漫画好きのユーザは、最近読んだ漫画の批評をブログ記事にすることが多い。また、趣味のブログ記事は一般的なブログ記事よりも専門性が増す傾向にある。そのため、野球好きな人は最近活躍しているプロ野球選手を他のユーザよりも先行して話題にしていると考えられる。また、漫画好きな人もまだ話題になっていない漫画の批評する可能性がある。

この観点から、共通の趣味を持つユーザ群をブログコミュニティとし、そのブログコミュニティ内での話題を抽出する。コミュニティ内では趣味のカテゴリを中心としたブログ記事が多く出現するようになり、コミュニティのカテゴリに特化した、今まで抽出できなかった話題の抽出ができると考えられる。

3.2 話題波及の検出

本研究では、ブログコミュニティのカテゴリを表1に示すように設定した。このカテゴリは、大カテゴリを12種類とし、小カテゴリを49種類とした。小カテゴリは1つの大カテゴリに属しており、表1のカテゴリは二階層のカテゴリと言える。このカテゴリは、yahoo! ブログ^{*}のカテゴリを参考に、作成した。

3.3 ブログ記事の分類

ブログコミュニティの分類は、ユーザの過去の記事を分類した結果を利用することで実現する。ブログ記事の分類を行うには、手動でカテゴリ分類したデータを使用する。このデータは、表1に示したカテゴリのタグが付いた複数のブログ記事である。タグ付けは、

^{*} <http://blogs.yahoo.co.jp/>

表 1. ブログコミュニティの分類に使用したカテゴリ.

大カテゴリ	小カテゴリ
政治経済ニュース	政治 経済 ニュース批評
スポーツ	野球 サッカー 格闘技 ゴルフ
音楽	邦楽 洋楽 ジャズ クラシック バンド
グルメ・フード	食べ歩き・外食 レシピ 料理
エンターテインメント	映画 テレビ 芸能人 芝居
芸術	文学 アート ファッション
乗物	車 電車 バイク 自転車 飛行機
ギャンブル	パチンコ・スロット 競馬 麻雀
生活	学校 家庭 恋愛 仕事 不安心理 インテリア 旅行 ダイエット健康
ペット・育成	犬 猫 その他の動物 ガーデニング
テクノロジー	コンピュータ インターネット 科学
遊び	おもちゃ 軍事 アニメ漫画 ゲーム

ブログ記事 1 件につき複数付けることが許される。例えば、「高校での野球部」を題材にしたブログ記事なら、タグは「学校」と「野球」の 2 つとなる。そのようなデータを利用して、以下の方法で記事の分類を実現する。

- (1) ブログ本文を形態素解析により、単語に分割する。
- (2) カテゴリごとの TFIDF 値を算出し、カテゴリを代表する単語を抽出する。
- (3) その単語が記事に含まれていたときに、記事が指定のカテゴリに分類される確率を計算する。
- (4) 単語とその確率を基に分類器を作成する。
- (5) 未分類ブログ記事の各カテゴリに属する尤度を算出する。
- (6) 各カテゴリの尤度を基に、未分類ブログ記事のカテゴリを判定する。

3.4 ブログコミュニティへの分類

ブログユーザのカテゴリは過去のブログ記事より判断できる。過去のブログ記事を分類したとき、あるカテゴリが付けられた記事数の割合が閾値を越えれば、ユーザもそのカテゴリに分類する。例えば、ユーザ u の過去の記事を分類し、「野球」と判断される記事の割合がある閾値を越えた場合、ユーザ u を「野球」のブログコミュニティに分類する。

3.5 話題の抽出

ここでは、話題抽出について説明する。話題が盛り上がるとは、内容が類似したブログ記事の数が一定期間で急激に増える状態を意味する。内容が類似した記事とは、本文中に登場する単語の分布が似ているかで判断する。一定期間で急激に増える状態の検出するために、バースト検出⁴⁾を参考にした。

4. 実験と結果

本システムの有効性を示すために実験を行った。

4.1 データセット

本システムで使用したデータセットの説明をする。使用したブログデータは情報処理学会 数理モデル化と問題解決研究会とソネットエンタテインメント株式会社において共同開催されたりコメンテーションサービスコンテスト⁵⁾より支給された。コンテストのデータは 2007 年 1 月 1 日から 2007 年 12 月 31 日まで投函されて日本語のブログデータである。ブログデータは 1 件のブログ記事を 1 つのインスタンスとして、パーナメントリンク、タイトル、日付、カテゴリ、ブログユーザの名前、記事のキーワード、ブログの URL、ブログのタイトルの属性が格納されている。

支給されてブログデータに、記事の本文は含まれていないため、本文を実験で使用する場合は URL を参照して取得することが求められた。本研究においてもパーナメントリンクからブログ記事にアクセスし、ブログ記事データの収集をした。さらに、取得したブログ記事本文を mecab⁵⁾により形態素解析した。形態素解析により得られた単語のうち、助動詞、助詞、句読点、記号を不要語として削除した。最終的には、パーナメントリンク、タイトル、日付、本文、本文の単語区切りを新たなブログデータとして格納した。

本システムを適用するために、ユーザごとにブログデータをまとめる必要があった。データは記事単位で格納されていたので、これをユーザごとにまとめた。ブログ記事の URL には、ユーザ名が埋め込まれていることが多い。これを利用し、ブログ URL からブログデータをユーザ単位に格納した。また、ブログ記事の HTML ファイルにある過去のブログ記事 URL を使用して、支給されたデータに存在しない記事の取得を行った。その結果、2004 年～2008 年のブログを収録することができた。

⁵⁾ <http://www.so-net.ne.jp/web2/compe2008/contest.html>

表 2 学習により得られた特徴単語 (抜粋)

カテゴリー	特徴単語
バンド	ライブ バンド ギター 練習 ACIDS 演奏 …
映画	観る 見る 映画 作品 監督 作 …
野球	試合 点 回 勝 移籍 戦 大リーグ …
パソコン	PC XP 修理 HDD パチ GB パソコン …
家庭	日 行く 笑う 言う 買う やる …

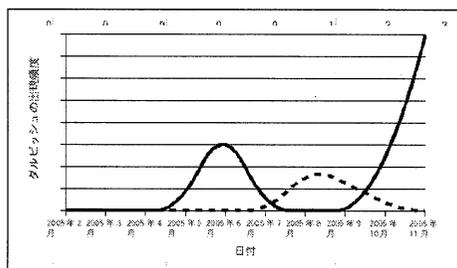


図 2 「野球」と「ゲーム」のブログコミュニティに出現する「ダルビッシュ」の出現頻度

データセットのサイズは、ユーザ数が 6,024 人、ブログ記事の総計は約 300 万件となった。また、このデータのうち 2,000 件を手動で表 1 に示すカテゴリに分類をし、それらをトレーニングデータとして扱った。

4.2 ブログコミュニティの生成

本システムを使用して、ユーザをブログコミュニティに分類した。

トレーニングデータを使用して、カテゴリを代表する単語の抽出を行った結果、表 2 ような単語が抽出された。比較的学習はうまく行えていることが分かる。しかしながら、「家庭」は特徴を示した単語とは言えない。これは、どのカテゴリにも属さない日々の身近に起こった出来事が書かれているブログのカテゴリを「家庭」と判断したため、あまり特徴的な単語を抽出できなかったと考えられる。

4.3 ブログコミュニティ内の話題抽出

システムが作成したブログコミュニティで、月ごとの単語の出現頻度を確認する。図 2 に、結果の一部をのせる。これは、「野球」と「ゲーム」のコミュニティにおける「ダルビッシュ」という単語の出現頻度を示したものである。

この図は、実線が「野球」のブログコミュニティの出現頻度の移り変わりを示していて、点線が「ゲーム」のブログコミュニティの出現頻度の移り変わりを示している。

この図から、「野球」のブログコミュニティでは「ゲーム」のブログコミュニティに先駆けて、「ダルビッシュ」の話題が盛り上がったことが分かる。また、その後「ゲーム」のブログコミュニティにも「ダルビッシュ」

の話題が盛り上がる時期が存在する。これにより、「ダルビッシュ」の話題は、「野球」から「ゲーム」に波及したと言える。

5. おわりに

本稿では、ブログコミュニティを形成し、全体に波及する前に話題を抽出する手法を提案した。実験では、ブログコミュニティにより話題の盛り上がり方が違うことを確認した。また、他のコミュニティであまり盛り上がりを見せていない時期にも、カテゴリに特化したコミュニティでは先行して話題が盛り上がることを確認した。これらのカテゴリに特化したコミュニティの話題を紹介することは有効であり、本システムの有効性が立証できた。

しかしながら、今回の研究では、どのようなときに話題が波及するのかまでは分からなかった。今後は、単独のコミュニティでの話題を抽出したあと、その話題が波及するかを判定する手法の開発をしていきたい。

参考文献

- 1) 内田誠, 柴田尚樹: ブログ記事ネットワークからの emerging topic の抽出と可視化, *Annual Conference of the Japanese Society for Artificial Intelligence*, Vol. 3D2, No. 03 (2006).
- 2) 南野 朋之奥村 学, 藤木稔明, 鈴木泰裕: blog ページの自動収集と監視に基づくテキストマイニング, *人工知能学会研究会* (2003).
- 3) 石田和成: 潜在的ウェブログコミュニティ抽出のための二部グラム分割アルゴリズム, *人工知能学会研究会*, pp. 1235-1244 (2004).
- 4) 岩井原瑞穂野真太郎, 階層型カテゴリを用いたウェブサイトのアクセス履歴の時系列相関性解析, *電子情報通信学会第 16 回データ工学ワークショップ*, No. 16 (2005).
- 5) 工藤拓: 形態素周辺確率を用いた分かち書きの一般化とその応用, *言語処理学会全国大会* (2005).