ブログのカテゴリ分類に基づく注目話題の早期検出

Early Topic Detection Based on Blog Categorization

小阪　有平[*1]　　安村　禎明[*1]　　上原　邦昭[*1]
Yuhei kosaka　　Yoshiaki Yasumura　　Kuniaki Uehara

[*1]神戸大学大学院工学研究科情報知能学専攻
Dept. of Information Science and Systems Engineering, Graduate School of Engineering, Kobe University

This report presents a method for early topic detection from blog articles. Existing methods for topic detection are usually based on burst detection. However, most bursted topics are already popular ones. The topics are not valuable information from the viewpoint of marketing. Valuable topics are described in a few blogs and have a potential to spread to many blogs. In this report, we propose a method for detecting potential topics from blogs based on blog categorization and machine learning. First, we collect blog articles and the blogers are classified into one category according to blog articles. The system extracts remarkable topics in the blog community (categorized blogers). Then we filter the topics based on its Document Frequency (DF). Next, we create a classifier for deciding a potential topics from the topic frequency transition in the blog community and all blogs. Finally, the system recommends the topics that decided on potentail topics by the classifier. Experimental results using actuall blog data show that the precision is 78.4% and the recall is 83.4% in potential topics detection. The results indicate that our method is effective for early topic detection.

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，ブログ記事がブロガの興味・関心をリアルタイムに反映することを意味する．このため，ブログ記事に記述された情報や話題を検出することは，トレンド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り [Uchida 06]，ブログからの注目話題を検出する手法が研究されている [Okumura 04, Ishida 03]．

ブログから注目話題を検出・検索するシステムは，現在多数稼働している．ブログ専用の検索システムでは Technorati[*1] や kizasi.jp[*2]などがあり，yahoo[*3]，google[*4]にもブログ検索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[*5]．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抽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．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．

いずれのシステムも話題検出には，バースト検出が用いられている．バースト検出はブログ全体で急激に増加し，バースト状態にある話題を検出することを指す．このよう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，マーケティングの観点から見ると，新たな需要を喚起するような情報にならない場合が多い．マーケティングなどに有用なのは，話題を早期に発見し，世間にさきがけて話題を提供することである．

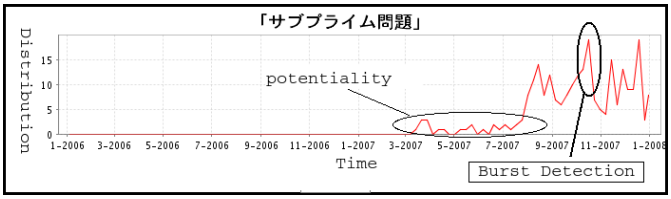

図 1: 早期話題として全体に先駆けて出現する話題の例

そこで，本研究ではブログのカテゴリ分類と機械学習に基づき注目話題をバースト以前に発見する手法を提案する．本手法ではブログをカテゴリに分類することによってその分野の専門家集団を抽出する．専門家集団とその他のブロガにおける話題の頻度推移に基づき機械学習によって注目話題を検出する．

2.　注目話題の早期発見

ここでは，注目話題の早期発見手法について述べる．まず，本手法の概要を説明する．次に，ブログのカテゴリ分類手法と機械学習による注目話題の検出手法を詳述する．

2.1　注目話題の早期発見の概要

需要を喚起する情報を発見するためには，話題を早期発見することが重要である．話題には，噂を呼ぶように徐々に大きくなり，最終的にバースト状態になる場合がある．その例を図 1 に示す．これは話題「サブプライム問題」に関する頻度推移である．この話題は 3 月にブログ記事に記述されるようになり，初めは極少数の頻度で記事にされる程度であったが，時間経過とともに頻度が増加していき，2007 年 10 月にバーストしている様子を示している．

また，図 2 に「サブプライム問題」の頻度推移を経済関連をブログ記事に書くことが多い傾向にあるブロガとそれ以外のブロガの頻度推移を示す．グラフから早期に記述しているのは，経済についてのブロガが大半を占めていることがわかる．

連絡先: 安村　禎明，神戸大学大学院工学研究科情報知能学専攻，〒 657-8501 神戸市灘区六甲台町 1-1　，TEL(078)803-6227，FAX(078)803-6316，yasumura@ai.cs.kobe-u.ac.jp

∗1　http://www.technorati.jp
∗2　http://kizasi.jp
∗3　http://blog.livedoor.com
∗4　http://blogsearch.google.co.jp
∗5　http:/blog.fkoji.com/2007/03270011.html

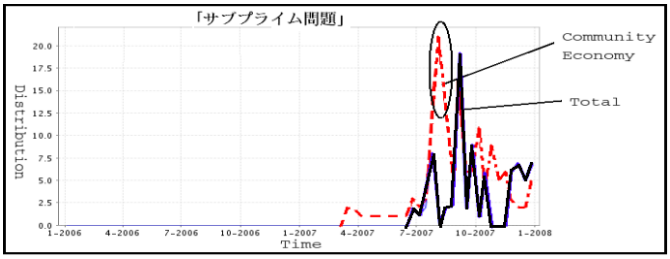

図 2: ブログコミュニティでの早期話題の推移の例

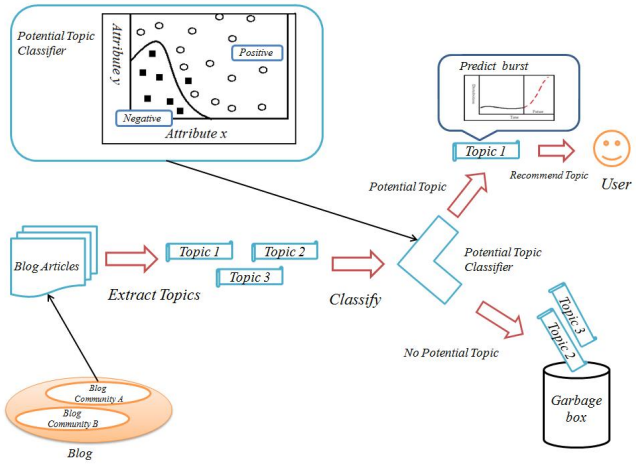

図 3: 提案手法の概要

そのため，話題を早期発見するためには，ブログの一部で盛り上がりを見せている話題に着目し，それらの話題が全体でバーストするかを判定することで実現できると考えられる．

以上のような発想をもとに，本手法ではブログのカテゴリ分類に基づく機械学習による注目話題の検出手法を提案する．本手法の概略を図 3 に示す．まず，ブログの一部における盛り上がりを得るためにブロガを専門性の高いコミュニティに分類する．コミュニティから話題の抽出を行うことで早期話題の候補とする．このとき，コミュニティで検出した話題の中には周期性のある話題（クリスマスなど）などが含まれるが，注目話題としては不適切である．早期話題の候補を絞りこむために Document Frequency（DF）値によるフィルターを生成し，話題を絞りこむ．

早期話題の候補とした話題の中で，ブログ全体も考慮した話題出現頻度の推移から学習することで，その話題が全体に波及するものか，全体に波及しないものかの判定を行う分類器の生成を行う．最後に，そこで作成した分類器を用いて，現在の話題を判別することで注目話題の早期発見を行う．

2.2 ブログのカテゴリ分類

ここでは，ブログカテゴリの詳細とその分類手法について述べる．

2.2.1 ユーザの専門性

ブロガーは，ブログ記事を自由意思のもとで作成しているが，その内容はブロガ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．このため，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を批評する可能性がある．

この観点から，共通の趣味を持つブロガをカテゴリ分類し，そのブログカテゴリ内での話題を抽出する．カテゴリ内では趣味の話題を中心としたブログ記事が多く出現するようになり，カテゴリに特化した話題を全体より先駆けて検出できることが期待できる．

2.2.2 ブログのカテゴリ分類

ブログのカテゴリ分類は，特定の趣味を持つブロガをその趣味に応じて分類することで実現する．ブロガの分類には，ブロガの過去の記事を分類した結果を利用する．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，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．こ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する．
2. カテゴリごとの TF*IDF 値を算出し，カテゴリを代表する単語を抽出する．
3. カテゴリを代表する単語をもとにナイーブベイズで，ユーザの未分類ブログ記事を分類する．

過去のブログ記事のカテゴリの中で記事数の割合が閾値を越えれば，ブロガもそのカテゴリに分類する．例えば，ブロガ 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，ブロガ u を「野球」のブログカテゴリに分類する．

2.3 一般語フィルター

本研究では，名詞および名詞連続に話題語の候補をしぼる．しかし，これではまだ早期話題の候補としては不適切である．本研究では注目話題の早期発見において効率良く学習を行うために，既知の話題をあらかじめ除去する．図 4 に「花火大会」のブログの話題推移の様子を示す．この「花火大会」は，夏期に盛り上がりを見せる話題であることが確認できる．このように，周期的にブログで盛り上がりを見せる話題が存在する．この「花火大会」の話題が高まる時期が分かったとしても，トレンド発見やマーケティングの観点から価値は低いと思われる．本研究では，既知の話題を提供することを目標としていないので，周期的に盛り上がりを見せる話題は，あらかじめ除去する．このような話題は一般語であるので，ブログの過去データから，各話題の DF 値を算出して，閾値を越える話題を削除対象にすることで実現できる．この一般語を除去することで，「オシムジャパン」「サブプライム問題」など未知の話題のみが残る．

2.4 注目話題分類器

一般語フィルタで，一般語を除去をしても，ここでは早期話題の他に，低頻度語が残る．低頻度語とはコミュニティや特定のブロガでしか語られない言葉のことである．「米連邦公開市場委員会」や「XX 町運動会」がこれに相当する．これらは DF 値が低いため，一般語の除去では取り除けない．このため，これら低頻度語と早期話題を判別する分類器が必要である．

表 1: ブログカテゴリ一覧.

Parent Category(大カテゴリ)	Son Category(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

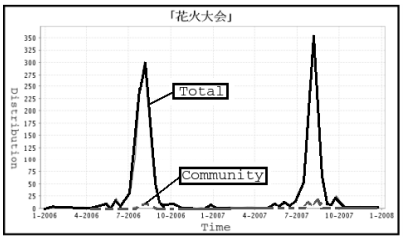

図 4: 花火大会の出現頻度推移

この分類器は話題の頻度推移から注目話題と低頻度語を分類する．注目話題には以下のような性質があることが多いため，この分類において考慮する必要がある．

- ブログカテゴリで話題がある程度大きくなっていること
- カテゴリ内だけでなく，ブログ全体で見ても話題に登りつつあること
- 話題が日を追うごとに大きくなっていること
- 連続的に話題に登っていること
- 話題がブログに初登場した時期からあまり経過していないこと

以上の事項を考慮し，話題の頻度推移から以下のような特徴を抽出し，分類器の入力とする．以下のすべての特徴量は数値属性をとる．

1. ブログカテゴリ内の当日，3 日前，7 日前，30 日前の話題頻度
2. ブログカテゴリ内の 1 日区切り，3 日区切り，7 日区切り，30 日区切りの連続登場回数
3. ブログカテゴリ内の初登場からの日数
4. ブログカテゴリで記事にしたブロガ数の累計
5. ブログカテゴリで過去 1 日，3 日，7 日，30 日で増えたブロガ数
6. ブログ全体の当日，3 日前，7 日前，30 日前の話題頻度
7. ブログ全体での 1 日区切り，3 日区切り，7 日区切り，30 日区切りの連続登場回数
8. ブログ全体で記事にしたブロガ数の累計
9. ブログ全体で過去 1 日，3 日，7 日，30 日で増えたブロガ数

分類器は，過去の事例をもとにした訓練データから機械学習を行うことで生成する．訓練データには，過去に一部で盛り上がりを見せ，その後全体に波及した話題を早期発見の正例とし，全体に波及しなかった例を負例とする．以上のような手続きによって注目話題の分類器を作成することで注目話題を検出する．

3. 実験

本手法の有効性を評価するために実際のブログデータを用いた実験を行った．

3.1 実験設定

実際のデータを使い早期話題を判別する分類器の性能を評価した．データには 2006 年 1 月〜2007 年 12 月の期間のブログデータを用いた．学習データは手動でタグづけした話題 300 件で，話題が最盛期になる時期の 3 日前，7 日前，30 日前を正例とした．負例は，話題候補の中で正例以外のものとした．ただし，話題が初登場した 30 日以内にバーストを起こしたものは，早期話題の定義である「ブログの一部で盛り上がりを見せている話題」を満たさないため，トレーニンデータに含めないこととした．実験では，学習器として C4.5 を使用し，10-fold cross validation による結果を求めた．

3.2 実験結果

実験結果として早期話題を検出した際の精度と再現率を求めた．精度とは早期話題と判定したもののうち，実際に早期話題であったものの割合であり，再現率とは実際の早期話題のう

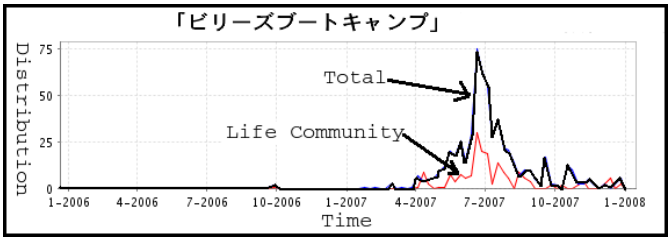

図 5: 抽出できた早期話題の例 1

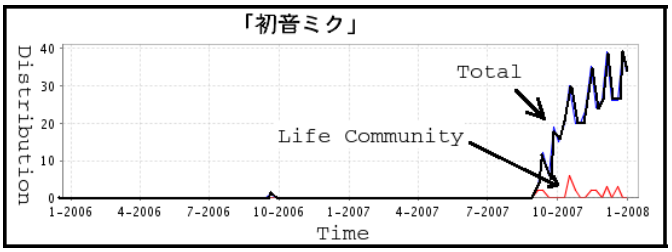

図 6: 抽出できた早期話題の例 2

ち，分類器のよって検出された割合を示す．この結果，精度が 78.4% であり，再現率が 83.4% であった．これらの結果は本手法が話題の早期発見に対して有効であることを示している．

本手法で，抽出できた早期話題とその後の推移を図 5，図 6 に示す．これらの話題はいずれも全体に波及している様子がわかる．一方，図 7，図 8 に早期話題として適切なものだったが，検出に失敗した例を示す．この例から出現頻度が急激に変化する場合は，検出に失敗することが分かる．また，図 9 に示すように，初登場と同時に話題が盛り上がるものについては，本手法を適用できない．このような話題はマスメディアによって取り上げられ，この影響によってバーストが発生したものである．

4. おわりに

本稿では，注目話題を早期発見する手法を提案した．早期発見を実現するために，ブログカテゴリとブログ全体での話題頻度推移を学習することで，話題が全体に波及することを判別する分類器を作成した．

検証実験から精度 78.4%で再現率 83.4%という高い検出結果が得られた．この結果から本手法はバースト検出以外の話題検出法を提案するとともに，ブログ全体に話題が波及する前に，話題を早期に検出することが可能であることを示した．

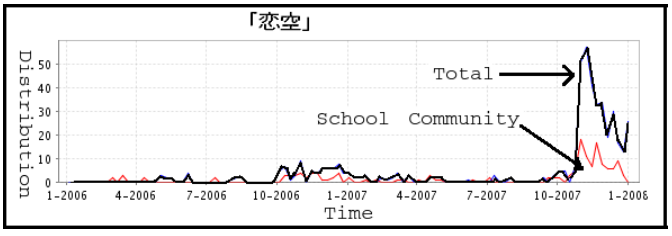

図 7: 抽出に失敗した例 1

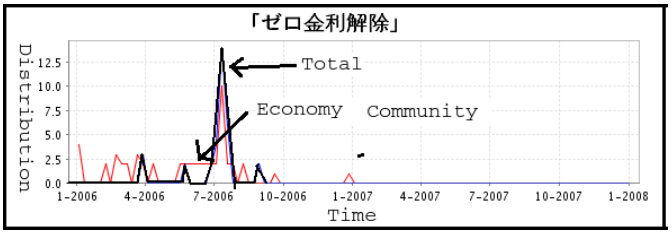

図 8: 抽出に失敗した例 2

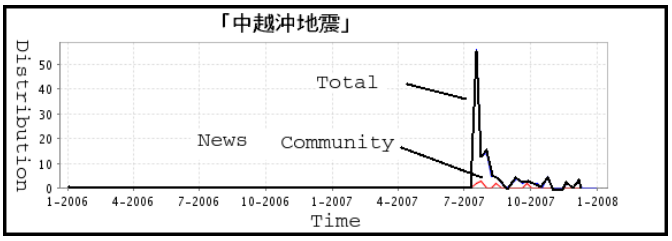

図 9: 本手法を適応できない例

また，取り出された話題推移を確認した結果，本手法は徐々に大きくなる話題を抽出することが可能であることが判った．

しかしながら，本手法で全体波及するを判定する分類器を作成するためには，大量のトレーニングデータを必要とし，これらの生成を手動で行うにはコストがかかる．分類器の作成には，少量のトレーニングデータから半教師付き学習を行う手法の開発が必要である．

参考文献

[Uchida 06] 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，人工知能学会全国大会講演論文集，Vol.20,No.3D2-03(2006)

[Okumura 04] 奥村学，南野朋之，藤木穂明，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資料，Vol.A401，No.01(2004)

[Ishida 03] 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，Vol. A401，No. 01(2003)