3D2-03

ブログ記事ネットワークからのemerging topicの抽出と可視化

Extracting and Visuzalization of a Emerging Topic from the Blogspace

内田 誠[*1] 柴田 尚樹[*2]
Makoto UCHIDA Naoki SHIBATA

[*1]東京大学人工物工学研究センター
Research into Artifacts, Center for Engineering, The University of Tokyo

[*2]東京大学大学院工学系研究科
Graduate School of Engineering, the University of Tokyo

Weblogs form a kind of network, whose nodes are their entries and edges are trackbacks between them. In this paper, analyzing the network, we propose a method to identify topics which are being discussed in individual entries, and to detect an emerging topic, on which many weblogs are vigorously discussiong at a moment. The method is based on network clustering and TF-IDF scoring on the contents of whole and clustered entries. Through an empirical study on a real dataset, we show that the weblog entries are effectively clustered according to their the topics, and, taking notice on a highly growing cluster, the emerging topic can be detected.

1. 序論

Weblogは，個人が手軽に情報を発信できるメディアとしての発展が著しく，Weblogを通じて発信される情報の増加や，そこでなされる議論の社会に対する影響の強さはもはや無視できない．Weblogの記事は日々生成され，蓄積されるという性質を有し，その全体は時系列的に動的に変化する．このことは，Weblogのコンテンツが書き手の興味・関心をリアルタイムに反映することを意味し，Weblog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を考える際に有用である．

Web上の記事群から特定の情報を見つけ出すことを計算機によって支援する方法には，主として記事中の文章を自然言語分析の方法によるものと，ハイパーリンクを分析するものの2つに大別される．前者の中では特にWeblogに特化して話題のトレンドを収集する手法が研究されている [Glance 04, 奥村 04, Gruhl 04]．また，Weblog記事に言及されるキーワードを集計し，その時点に多く記事で言及されているキーワードをランキング形式で提示するWEBサービスもいくつか実用化されている[*1]．一方，ハイパーリンクを分析する手法としては，Weblog記事の接続関係からコミュニティ構造を発見するための研究が主として行われている [谷口 04, 石田 04, Kumar]．特に，Weblogシステムには，記事同士の関連を明示するためのTrackbackの仕組みが実装されており，記事とTrackbackによってネットワーク構造が形成されていると見なすことができる．複雑かつ大規模なネットワークを取り扱ってその構造を分析する方法は，近年盛んに研究が行われている領域でもある [Newman 03]．

これらの方法には，それぞれ短所長所が存在する．記事中の語の頻度を解析する方法の場合，どの語がどの時期に頻繁に利用されたかは解析可能であるが，それらの語が言及された個々の記事がいつ，どのように出現し，成長したかを分析することや，それらの記事の関係性を把握することが難しい．他方，ハイパーリンクを分析する方法は，密に結合する記事群やその発展の過程を特定することは可能であるものの，それらの記事群がどのような話題を扱っているのかは特定できない．

本研究ではこれらの2つの手法を組み合わることによりWeblogの記事空間を分析し，記事群が話題にしているトピックを同定するとともに，各時点で急激に盛り上がるトピック (emerging topic) を抽出する方法を提案する．具体的には，Trackbackの仕組みによるネットワーク性に着目し，記事とトラックバックからなるネットワークに対して，ネットワーク分析によるクラスタリングと，自然言語処理による特徴語抽出を組み合わせ，ネットワークの時系列的成長と併せて分析することで，ある時点で急激に成長する記事クラスタと，そのクラスタの扱うトピックを特徴語の形で特定する．

2. 手法

2.1 記事の収集

本研究では，Weblogの記事をノード，Trackbackをエッジと見なし，図1に示したように各記事へTrackback元の記事を辿ることでWeblog記事のネットワークを取得する．起点となる記事は，Trackbackを複数獲得している記事を無作為に1つ抽出し，そこからTrackbackが途切れるまでクロールを繰り返した．さらに，各記事が生成された日付を取得する[*2]．なお，記事の収集は2004年12月に行った．

2.2 クラスタリング

次に，得られたWeblog記事のネットワークをネットワークトポロジによるクラスタリング手法によって分割する．これは，図2のように全体に比べてエッジが相対的に密であるような記事群に分割する方法であり，多変量解析で用いられるクラスタリングのように何らかの属性変数を元にするものではない．クラスタリングにはNewman[Newman 04]によって提案された方法を用いる．

2.3 クラスタ特徴語の抽出

2.2項で得られたクラスタに含まれる各記事の本文の特徴語を抽出するとともに，以下の手順で各記事の特徴語を集計することで，「クラスタの特徴語」を抽出する．

連絡先: 内田 誠 (uchida@race.u-tokyo.ac.jp)
東京大学人工物工学研究センター
千葉県柏市柏の葉5-1-5

*1 たとえば，http://kizasi.jp/やhttp://blog360.jp/などがサービスの一例である．

*2 生成日は個々の記事の本文から日付のフォーマットをパースする方法で抽出することを試みた．

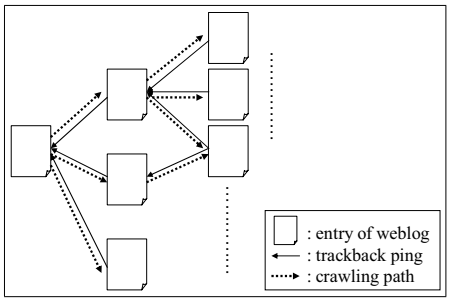

図 1: Weblog 記事の Trackback ネットワークとトレース方法.

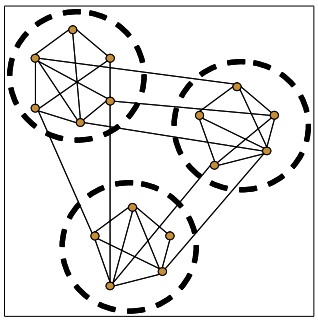

図 2: ネットワーククラスタリングの概念図.

1. 以下の手順で各記事の特徴語を 10 語を計算する.
   (a) 各記事の本文から HTML タグを除去する.
   (b) bulkfeeds[*3]の API を用いて, *tf-idf* が高い 10 語を抽出する[*4]. この 10 語の当該記事の特徴語とする.
2. 各クラスタの特徴語 10 語を, 語 i のクラスタ j に含まれる記事群における *tf-idf* によって計算する.

ここで抽出したクラスタ特徴語を, クラスタの時系列的成長と重ね合わせることで, ある時点で急激に成長しているクラスタと, そこで扱われている話題を特徴語の形で抽出することができる.

3. 結果と考察

3.1 クラスタの抽出と特徴語

2.1 の方法によって記事を収集した結果, 39,408 記事が得られ, そのうち生成日を正しく取得できたものが 25,668 記事であった. また, Trackback は送信側と受信側で向きのあるリンク情報であるが, 本研究では無向のエッジであると見なした. エッジ数は 60,892 本であった.

この記事ネットワークに対して, Newman 法によるクラスタリングを行った結果, 127 クラスタに分割され, $Q_{max} = 0.94$ であった. Q 値は分割されたクラスタ分割の明瞭さを表し, 0 から 1 の値を取る. Weblog 記事ネットワークの Q 値は 1 に近く, 非常に明瞭なクラスタ構造を有していることがわかる.

クラスタのうち, 含まれる記事数が 500 を超えるものの特徴語を表 1 にまとめる. また, ネットワークを描画し, 特徴語を付与した可視図を図 3 に示す. 図 3 では, エッジのみを描画し, 属するクラスタによって色分けしている. ノードの配置計算および描画には LGL [*5] を用いた.

クラスタ特徴語は定性的にはおおむねそのクラスタの記事の内容を代表しているように見えるが, ID が 4, 5, 7 のクラスタのように一部のクラスタではノイズと思われる語が多く混ざっており, 特徴語からは内容を特定しにくい. これは, 現時点では記事本文のみを抽出することが難しく, リンクやナビゲーションを含む記事全体を解析対象にしていることや, 辞書が不完全で一部の固有名詞には対応できないことが原因として考えられる. 実際に, ID4 や ID7 のクラスタには, 季節をテーマとした写真を共有する Weblog が多く含まれ, 相対的に本文の文章量が少ない傾向がみられた.

3.2 時系列成長とその可視化

得られたクラスタのうち, 一例として特に記事数の多い ID1 から ID3 のクラスタについて, 含まれる記事の生成数を日付別にプロットしたものを図 4 に示す. クラスタ特徴語から, これらのクラスタの代表的なトピックはそれぞれ「新潟中越地震」「プロ野球新規参入問題」「年金制度問題」をトピックとしてとして特徴づけることができる.

「新潟中越地震」クラスタでは 2004 年 10 月 20 日前後,「プロ野球新規参入問題」クラスタでは 2004 年 9 月 15 日前後に急激に新着記事数が増加している様子が見られる一方で,「年金問題」クラスタでは, 明確なピークがみられない. このように, トピックの性質によって Weblog 上での話題の盛り上がりには差異がみられる. 対応する出来事として, 2004 年 10 月 23 日に新潟県中越地震が発生した. また 2004 年 9 月 17 日にはプロ野球で史上初のストライキが決行された. これらの出来事をトリガーとしてトピックが急激に盛り上がった様子が観察される. 一方で,「年金制度問題」は特定の出来事に呼応するというよりはむしろ 2004 年を通じて長く議論されたトピックでもあり, その特徴が記事のクラスタの成長にも現れている.

また, 2004 年 10 月 23 日とその 1 週間後の Weblog 記事ネットワークの様子を描画したものを図 5 に示す.「新潟中越地震」クラスタが, 明らかにこの 1 週間のうちに急激に成長している様子が見られる. このように, Weblog 記事ネットワークのクラスタリングとその成長の様子から, その時点の emerging-topic とそれを扱う記事の集合を抽出することができる.

しかしながら,「新潟中越地震」クラスタの記事の増加には他にも複数のピークがみられる. たとえば, 2004 年 11 月 2 日には最も高いピークがみられるが, この日は東北楽天ゴールデンイーグルスのプロ野球への新規参入が決定した日である. 表 1 にも一部がみられるが, このクラスタは新潟中越地震に関連した記事に加え, プロ野球新規参入問題に関連した記事が一部混在していることがわかる. このように, Newman 法によるクラスタリングでは, ネットワーク構造の情報のみを利用するため, 記事が扱うトピックを完全には分離できていない様子が見られる. よりクラスタ抽出の精度を上げるためには, ネットワーク構造によるクラスタリングに加え, 記事の特徴語による分類などを併用する必要があると考えられる.

4. 結論と展望

本研究では, Weblog 記事が持つ Trackback による連結性に着目し, ネットワーク分析によるクラスタリングの手法によ

*3 http://bulkfeeds.net/

*4 bulkfeeds では, ping サーバに送られた直近の新着 100000 記事から *idf* を作成して, 記事の単語の *tf* を掛け合わせる *tf-idf* を用いている. 日本語の形態素解析には kakasi を用いている.

*5 http://apropos.icmb.utexas.edu/lgl/

表 1: 記事クラスタの特徴語（記事数の多いクラスタを抜粋）.

ID（順位）	記事数	特徴語（上位 10 語）
1	2934	地震，中越，新潟県，新潟，楽天，被災者，被災地，災害，募金，ライブドア
2	2409	野球，ライブドア，プロ，楽天，球団，9，スト，|，イチロー，ストライキ
3	1164	木村，剛，年金，週刊，ココログ，公的，トラックバック，ウェブログ，ブログ，チチ
4	1123	秋，@，September，甲信越，atom，yes!，長月，祭り，74，琥珀色
5	1102	玉，goo，Ken，2004，スノー，コメント，2005，トラックバック，戦隊，年
6	938	SEO，コンテスト，キーワード，第，検索，トラックバック，トラックバックセンター，google，Google，続き
7	820	夏，@，REI，8 月 11 日，Main，Title，Search，夏物，カーテンコール，! @
8	715	売却，球団，西武，ダイエー，ライオンズ，オリックス，野球，楽天，ホークス，ドラフト
9	668	スト，合併，ストライキ，プロ，野球，球団，近鉄，9，回避，オーナー
10	620	出会い，攻略，系，記事，|，アダルト，エログ，URL，エッチ，オナニー
11	617	@，秋，September，甲信越，atom，長月，yes!，74，祭り，琥珀色
12	563	近鉄，合併，野球，ナベツネ，オーナー，7，球団，買収，リーグ，プロ

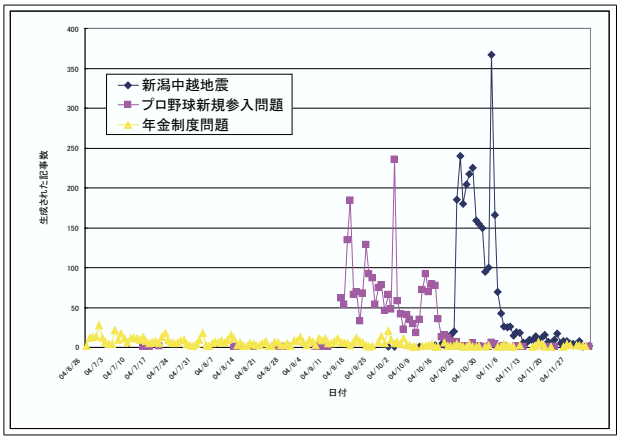

図 4: クラスタの記事の日付別生成数（記事数上位 3 クラスタ）.

り Weblog 記事の集合から扱う話題ごとの記事群に分割する方法を提案した．また，分割された記事群が取り扱う話題を特定するために，クラスタ内の記事の *tf-idf* によってクラスタの特徴語を抽出する方法を提案した．その結果，記事本文の情報を利用せずに Trackback のトポロジ構造によるクラスタ分割だけでも，話題ごとの記事群を同定し，そこに含まれる記事が扱う話題をクラスタ特徴語によって粗視的に把握することができることを示した．

また，Weblog 記事ネットワークの時系列的な成長と，分割されたクラスタの成長を重ね合わせることで，ある時期に急激に成長している記事のクラスタを特定し，その特徴語を調べることで，その時点で盛り上がっている話題とその記事の成長を特定する方法を提案した．その結果，Weblog 記事ネットワークにおけるトポロジ的なクラスタは，現実の出来事に呼応して急激に成長する時期を持つことがあり，ある時期で急激に成長しているクラスタとその特徴語に着目することで，その時点で盛り上がっている話題（emerging topic）を，それに言及する記事とその関係性を含めて抽出できることを示した．

今後の課題としては，本文中でも述べたように，ネットワーク構造によるクラスタリングに加え，記事内容によるクラスタリングを補助的に用いることで，クラスタ分割の精度を向上させる余地があると考えられる．また，本研究で用いた Newman 法では，トポロジの情報のみでクラスタリングを行うことが可能であるものの，分割のためにネットワーク全体の情報を必要とする．そのため，リアルタイムに成長を続けるネットワークの「現在時点」のクラスタの成長の様子を知るためには直近 2 時点での様子を比較する必要があるほか，記事が蓄えられデータが多くなると計算コストも増大する．このため，大規模かつ日々成長する対象に適した新たなネットワーククラスタリングの手法を開発することも課題である．

参考文献

[Glance 04] Glance, N. S., Hurst, M., and Tomokiyo, T.: BlogPulse: Automated Trend Discovery for Weblogs, in *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics* (2004)

[Gruhl 04] Gruhl, D., R, G., Liben-Nowell, D., and Tomkins, A.: Information Diffusion Through Blogspace, in *Proceedings of the 12th international conference on World Wide Web*, pp. 491 – 501 (2004)

[Kumar] Kumar, R., Novak, J., Raghavan, P., and Tomkins, A.: On the Bursty Evolution of Blogspace, *World Wide Web*

[Newman 03] Newman, M. E. J.: The Structure and function of complex networks, *SIAM Review*, Vol. 45, pp. 167 – 256 (2003)

[Newman 04] Newman, M. E. J.: Fast algorithm for detecting community structure in netwoks, *Physical Review E*, Vol. 69, No. 066133 (2004)

[奥村 04] 奥村 学, 南野 朋之, 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 SIG-SWO-A401-1 (2004)

[石田 04] 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フ分割アルゴリズム, 人工知能学会 SIG-SWO-A404-01 (2004)

[谷口 04] 谷口 智哉, 松尾 豊, 石塚 満：Blog コミュニティの抽出と分析, 人工知能学会 SIG-SWO-A401-8 (2004)

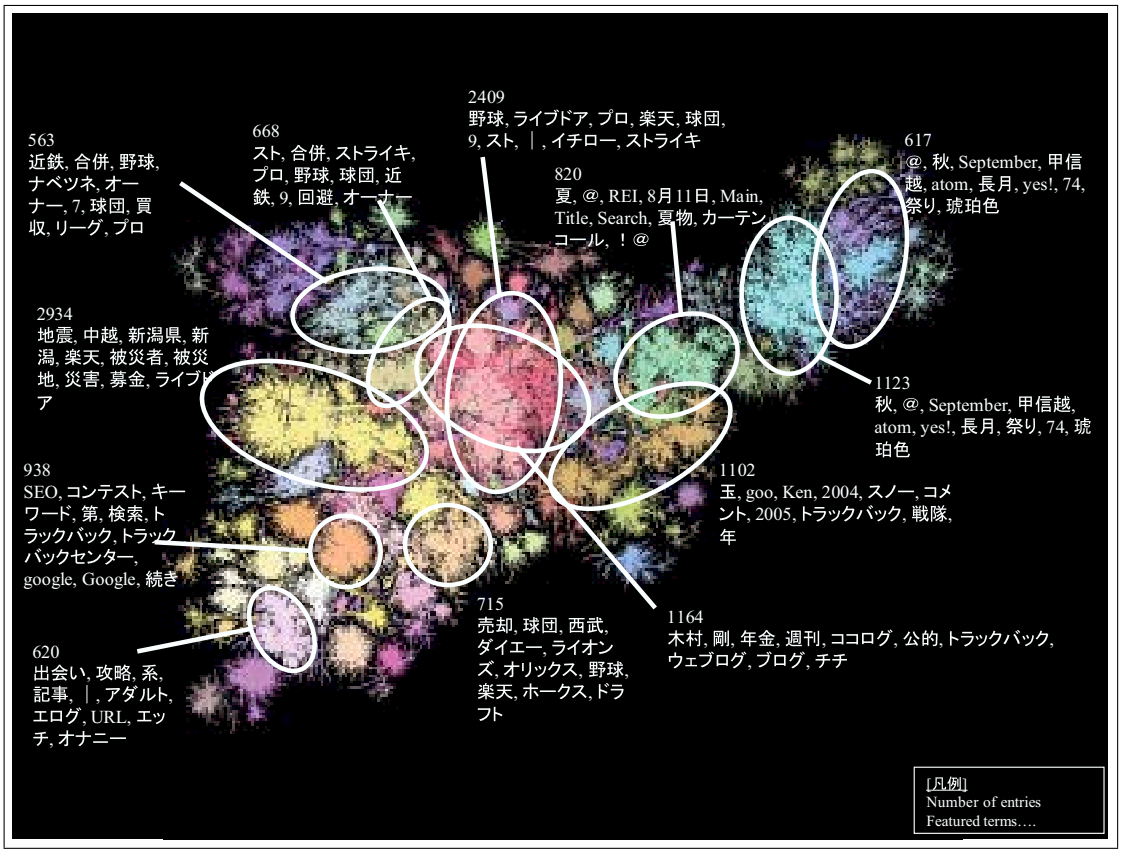

図 3: Weblog 記事ネットワークのクラスタ分割と特徴語.

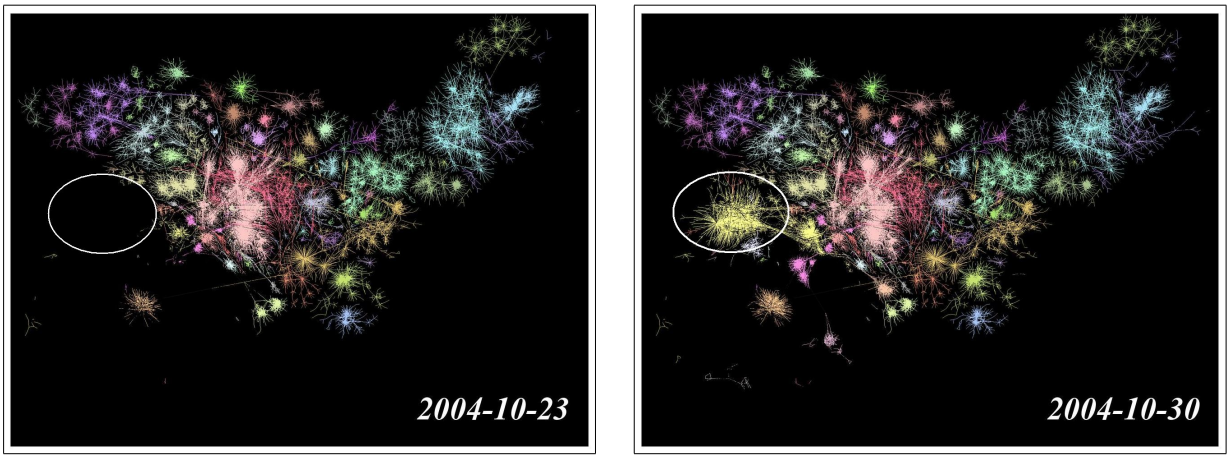

図 5: 「新潟中越地震」クラスタの成長（左： 2004 年 10 月 23 日，右： 2004 年 10 月 30 日）.