Blog における話題遷移点の検出

谷内　幸憲[†]　徳永　幸生[‡]　杉山　精[‡]

芝浦工業大学大学院電気電子情報工学専攻[†]　芝浦工業大学工学部情報工学科[‡]

1. 研究の背景・目的

近年，Blog はその利用の簡便さから急速に普及が進み，誰でも簡単に Web 上に情報発信できる時代になった．日本においても Blog の開設数は増加しており，総務省調査[1][2]では 2005 年 9 月末から 2006 年 3 月末までの半年間に約 533 万件ものBlog が新たに開設されている．

しかしその一方で，誰でも手軽に書けるが故に Blog 上に存在する情報の量は膨大なものとなり，人手でその中の情報を整理したり，有益な情報を探し出す事は困難になっている．

そこで本研究では，トラックバックを利用して Blog 間で展開される一連の話題（以下，Blog スレッド[3]）に注目し，その話において話題の変化のきっかけとなっているエントリー（以下，話題遷移点）を検出し，Blog 上の情報整理を試みる．

2. 提案手法

一般の会話や議論において話題が変化するような発言をした場合，会話の参加者はその発言に対して何らかの反応を行う．これを Blog における会話に適用すると，話題遷移点に対するトラックバックは閲覧者からの反応であると考えられる．そしてこの時，トラックバックを送ったエントリーでは新しい話題に対して言及をしている可能性が高いと考えた（図 1）．

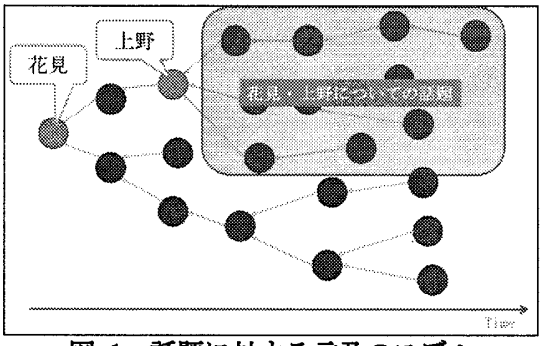

図 1：話題に対する言及のモデル

Detection of topic transition in Blogs
Yukinori Taniuchi†, Yukio Tokunaga‡, Kiyoshi Sugiyama‡
Graduate School of Engineering, Shibaura Institute of Technology†
Shibaura Institute of Technology‡

そこで，あるエントリーの前後のエントリー群で全ての単語について次の評価式を適用・比較し，そのエントリーへの言及を検出することで話題遷移点の検出を行う．

評価式：$F_{ev} = tf_i \cdot df_i$

tf_i：単語 i の出現頻度
df_i：単語 i を含む文書出現頻度

ここで，評価式の値が前後のエントリー群の間で最も大きく変化しているエントリーを話題遷移点であると仮定する．

3. システムの概要

3.1 記事データの収集

始めに，あらかじめ goo ブログの id を約 38 万件用意し，それらの中からランダムに 150 個の id を抽出した．次に抽出した id の blog エントリーをクローラを用いて全件取得し，更にその記事群に含まれるトラックバックを再帰的に取得した．

今回は記事データベースとして約 9 万エントリーを用意した．

3.2 Blog スレッドの抽出

3.1 で取得した各エントリーのトラックバックを辿りながらスレッド id を割り当て，Blog スレッドとして抽出する．

3.3 話題遷移点の検出

3.2 で抽出した Blog スレッドに対して提案手法を適用し，話題遷移点を検出する．本システムでは提案手法を次のように実装した．

まず始めに各エントリーの本文を抽出する．そして抽出した本文それぞれに対して MeCab[4]による形態素解析を行い，エントリー毎の単語の出現頻度を調べる．

次に各エントリーそれぞれについて，前後に接続されているエントリー群を Blog スレッドから抽出し，先に調べた単語出現頻度を用いて各単語の文書出現頻度を求め，評価式を適用する．そこでエントリー前後の評価値の差があらかじ

め定めた閾値以上となる単語があった場合，そのエントリーを話題遷移点とし，その単語を話題語とする．

4. 実験と結果

抽出した Blog スレッドの中でエントリー数が最も多かったスレッドを対象として話題遷移点の検出を行った．

一般の Blog を対象に本文やトラックバックを検出するためには各 Blog サービスに合わせた大量のテンプレートに対応する必要があるため，実験では検出の対象を goo ブログに限定した．また，各エントリー前後のエントリー群はステップ数 2 までのものとし，評価値の差の閾値は 1.0 とした．

また，対象のスレッドでは映画に関する話題が展開されており，図 2 の上側のエントリーでは「Shall we dance?」に関する話題，下側のエントリーでは「交渉人 真下正義」に関する話題が主に展開されていた．

そして，このスレッドの上側のエントリー群に対してシステムによる話題遷移点の検出を行って色を付け，左から順に時系列に並べたのが図 3 である．

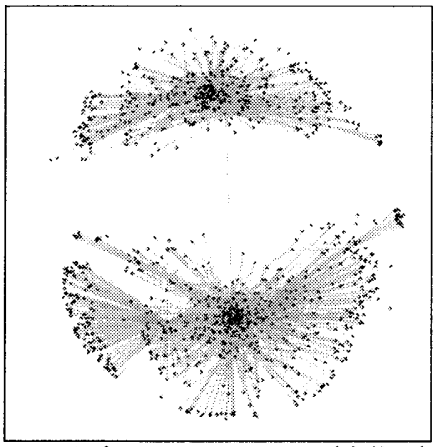

図 2：対象 Blog スレッド （全体図）

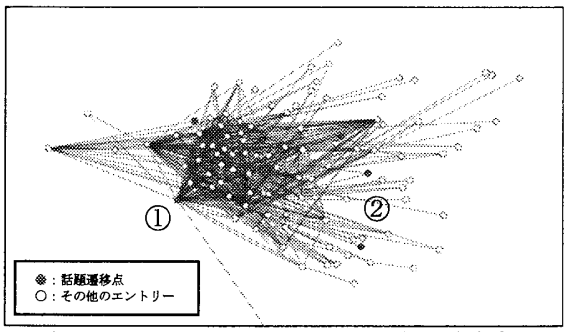

図 3：Blog スレッドとその話題遷移点

5. 考察

図 3 内の①の話題遷移点では話題語として「真下」が検出されている．このエントリーは図 2 の下側と接続している点であるので話題の変化を上手く捉える事が出来ていると言える．

また図 3 内の②付近の話題遷移点では「記事」，「男性」といった単語が話題語として検出されているが，これらのエントリーはまだトラックバックされていないため，今までのエントリーになかった傾向の単語が検出されていると考えられる．ここでの話題語は今後の話題の芽として考える事が出来る．

一方，本手法の問題点として，この部分での中心的な話題である「Shall we dance?」についての検出が上手くできていないという点がある．これについては，本手法では話題語の変化を検出しているため，一番始めから含まれる話題を検出できなくなっていると考えられる．この問題を解決するには，閾値の調整や一番始めの話題語については別の手法を用いるなどの対策が必要である．

6. まとめ

本研究では新しい話題に対する言及に着目し，単語出現頻度の変化による話題遷移点検出モデルを提案した．このモデルに基づき，実際に Blog スレッドに対する話題遷移点の検出を行ったところ，本モデルが有用であるとの見通しを得た．

今後は一番始めの話題検出に関して検出手法の改善を検討する．また，goo ブログ以外の Blog サービスへの対応も進めることで Blog スレッドの全体を対象と出来るようにし，正確に話題遷移の検出を行えるようにしたい．

参考文献

[1] ブログ及び SNS の登録者数（平成 17 年 9 月末現在），http://www.soumu.go.jp/s-news/2005/051019_2.html

[2] ブログ及び SNS の登録者数（平成 18 年 3 月末現在），http://www.soumu.go.jp/s-news/2006/060413_2.html

[3] 中島伸介，舘村純一，原良憲，田中克己，植村俊亮：“重要な blogger 発見を目的とした blog スレッド解析手法“，知能と情報(日本知能情報ファジィ学会誌)，Vol.19，No.2，pp.156-166，Apr.2007

[4] MeCab，http://mecab.sourceforge.net/