主観評価に基づくタグ重み付けによるblog話題遷移点の抽出

谷内　幸憲[†]　徳永　幸生[††]　杉山　精[††]　杉崎　正之[†††]　望月　崇由[†††]
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[††]
NTTレゾナント株式会社[†††]

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

そこで我々は，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[1]）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）を抽出することで，Blog上の情報整理を試みてきた[2]．この話題遷移点の抽出における重要な要素として，各単語の話題性の評価式がある．本稿ではこの話題性評価式にBlogエントリー中のHTMLタグによる重み付けを加える評価式を提案し，それによる話題遷移点抽出精度の改善効果を検証した．

2. 提案手法

会話や議論において話題が変化するような発言があった場合には，一般に会話の参加者からはその発言に対する何らかの反応がある．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）．

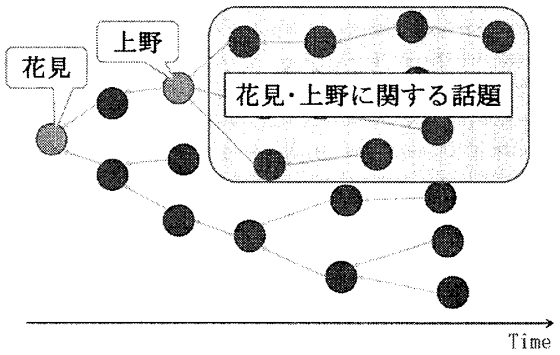

図1：話題に対する言及のモデル

そこで，各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(2.1)を適用し話題性の評価を行う．

評価式：$F(t_i) = tf_i \cdot (\{d : d \in t_i\}/D)$　　…(2.1)

tf_i：単語t_iの出現回数
d：単語t_iを含む文書出現回数
D：対象エントリー群の文書数

ここで，評価式(2.1)の値が前後のエントリー群の間で大きく変化しているエントリーを話題遷移点であると判定する．

本稿では，この手法による抽出精度を更に向上させるために，評価式(2.1)に対して以下のようにHTMLタグによる重み付け補正を行った．

評価式：$F(t_i) = tf_i \cdot (\{d : d \in t_i\}/D) \cdot \max(c(t_i))$　…(2.2)

$c(t_i)$：単語の含まれるタグの補正値

そして，このタグの補正値を決定するため，人の主観評価による話題抽出実験を行った．

3. 話題遷移点抽出システムの概要

3.1 記事データの収集

任意のBlogエントリーから本文中のリンクとトラックバックを辿って一連のBlogスレッドを取得するクローラーを作成した．

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，Blogスレッドとして抽出する．

3.3 話題遷移点の抽出

3.2で抽出したBlogスレッドに対して提案手法を適用し，話題遷移点を抽出するため，次のような操作を行う．

ステップ 1:あるエントリーの前後に接続されたエントリー群を取得する．

ステップ 2:前後のエントリー群それぞれにおける各単語の出現回数，出現エントリー数などを計算する．ここで本システムでは話題語の品詞を名詞に限定し，更にストップワードなどを用いて不要語の削除を行う．

ステップ 3:ステップ2で求めた値を用いて評価式を適用し，前後のエントリー群の評価値の差が一定値以上になった単語があれば，このエントリーはその単語を話題語とした話題遷移点であると判定する．

4. 実験と結果

タグの補正値を決定するため，実際のBlogスレッドから人の主観評価によって話題語を抽出する実験を行った．その結果，抽出された話題語の属する主なタグの出現率は次のようになった（表1）．

Approach to Detect Topic Transition in Blogs by Subjective Evaluation based on Tag Weight
Yukinori Taniuchi†, Yukio Tokunaga††, Kiyoshi Sugiyama††, Sugizaki Masayuki†††, Mochiduki Takayoshi†††
Graduate School of Engineering, Shibaura Institute of Technology†, Shibaura Institute of Technology††, NTT Resonant Inc.†††

表 1：話題語の属する主なタグ出現率

span	0.004587156	a	0.565749235
h1	0.003058104	p	0.318042813
h2	0.001529052	li	0.018348624
h3	0.027522936	blockquote	0.012232416
h4	0.012232416	b	0.003058104
div	0.027522936	strong	0.004587156
ins	0.001529052		

一方，話題語以外の語も含む全ての語の属する主なタグの出現率は次のようになった（表 2）．

表 2：全単語の属する主なタグ出現率

span	0.050140632	a	0.272014280
h1	0.000378624	p	0.458838165
h2	0.001081783	li	0.060038944
h3	0.008546084	blockquote	0.010168758
h4	0.007626569	b	0.000216357
div	0.064744699	strong	0.001947209
ins	0.000324535		

ここで表 1 と表 2 の各タグの出現率を比較すると次のようになる（図 2）．

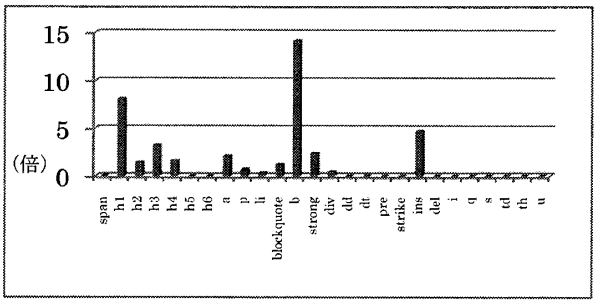

図 2：タグの出現率比較(表 1/表 2 倍)

そこで本システムではこれらのタグの出現率を用いて，各タグについて次の式(2.3)で求めた値を補正値とした．

$$\text{補正値 } c = \begin{cases} \ln(tf_t / tf_a) & \dots (tf_t / tf_a) \geqq 1 \\ 0 & \dots (tf_t / tf_a) < 1 \end{cases} \quad \cdots (2.3)$$

tf_t:話題語のみにおけるタグの出現率
tf_a:全ての語におけるタグの出現率

5. 評価と考察

従来の話題遷移点抽出システムと，4 の実験で得られたタグ補正値を用いた新システムのそれぞれで話題遷移点を抽出した結果，次のような特徴が見られた（図 3，図 4）．

(1)新システムでは話題遷移点の抽出数が増加し，今まで抽出できなかった話題を多数抽出できるようになった．しかし，旧システムよりも同じ話題が何度も抽出されてしまうケースが増加するという問題も発生した．

(2)話題語自体は抽出できているが，人の評価による話題遷移点よりも前後した位置で検出されている場合が新旧のシステムともに存在した．

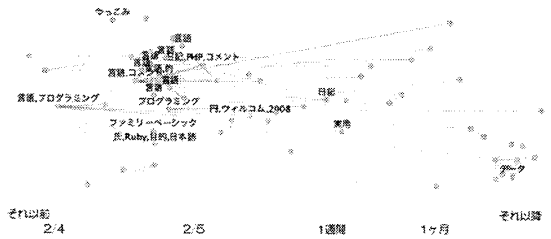

図 3：旧システムによる話題遷移点グラフ

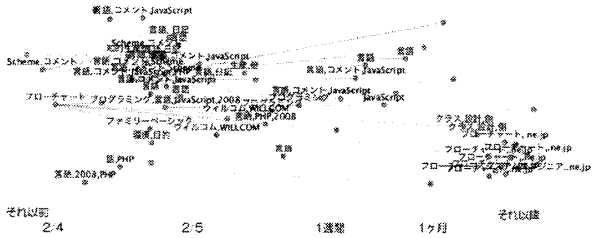

図 4：新システムによる話題遷移点グラフ

(1)については，タグによる補正が強く効き過ぎたか，従来の評価式(2.1)における話題の伝播性の評価項が今回の変更によって有効に機能しなくなっている可能性がある．従って，タグによる補正項の影響について更に詳しく調べる必要がある．

また，(2)については，現在の評価式(2.1，2.2)では各エントリー群内でのエントリー間の距離などを考慮していないため，このような微妙な話題遷移点のズレが起こってしまうと考えられる．評価式の形について距離などを考慮した新しいアプローチを模索する必要がある．

これらの問題点はあるものの，新システムは旧システムに比べて話題の抽出精度が向上しており，今回のタグを用いた補正は話題遷移点の抽出に対して有用であると言える．

6. まとめ

本稿では HTML のタグに重み付けを導入した話題遷移評価式を提案した．また，実際に人の主観評価による話題抽出実験を行い，その結果を用いて話題遷移点の検出を行ったところ，話題語抽出に関しては本手法が有用であるとの見通しを得た．

今後は話題遷移点の抽出に関して手法の改善を検討する．また，話題遷移点の提示手法に関しても研究することで Blog スレッドへの効率的なアクセスを実現していきたい．

参考文献

[1] 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

[2] 谷内幸憲，徳永幸生，杉山精：“Blog における話題遷移点の検出”，第 70 回情報処理学会全国大会，1-615，Mar. 2008