2C2-01

Weblogにおける語の広がり方に基づいたキーワード抽出

Keyword extraction based on a spread of the word in Weblog

岡田 健[*1] Takeshi OKADA
松澤 智史[*2] Tomofumi MATSUZAWA
松尾 豊[*4] Yutaka MATSUO
内山 幸樹[*3] Koki UCHIYAMA
武田 正之[*2] Masayuki TAKEDA

[*1]東京理科大学大学院 理工学研究科 情報科学専攻
Graduate School of Science and Technology, Tokyo University of Science

[*2]東京理科大学 理工学部 情報科学科
Tokyo University of Science

[*3]株式会社ホットリンク
hottolink,Inc.

[*4]独立行政法人産業技術総合研究所 サイバーアシスト研究センター
National Institute of Advanced Industrial Science anc Technology

In recent years, the individual is sending new information one after another. This means the sources of information in which we get quickly the information about the new occurrence and subject of a world. In order to employ the feature of Weblog as such media efficiently, it becomes important how a new topic can be taken out. In this research, the technique of extracting the word which is in fashion in it for Weblog, and subject is proposed. By the proposal technique, how to depend on the word which is in fashion on Weblog, and subject spread is analyzed and used. Furthermore, the technique of putting up a word in fashion and a topic is built through analysis of process in which this word and topic are in fashion. Moreover, the evaluation experiment of the proposal technique was conducted and the validity of the proposal technique was evaluated.

1. はじめに

インターネット上の不特定多数に情報発信する手段として、近年、ウェブログが増えつつあり、略してブログと呼ばれている [1]。そしてブログには、「Doblog」[*1]や「ココログ」[*2]などの無料のレンタルサイト、ウェブログの生成・管理を行う多機能なツールである「Blogger」[*3]や「Movable Type」[*4]などのツールが存在する。

これらのブログは、個人の興味や関心を発信するという点で優れている。ブログは、一般的に時系列に個人が論評した記事を記録しているサイトと定義されるように、時系列に記録された記事が個人の存在を示すことができる。さらに記事を閲覧した個人からのフィードバックを自動的に反映される相互参照が盛んな特徴もある。こうした相互参照性が強さは、自然と興味ある話題が「口コミ」のように広まるメカニズムが働くことを意味し、このような広まり方が絶えず新しい情報を生み出す循環を作っている。

またブログには、個人の更新が早いという特徴があり、これは世の中の新しい出来事や話題についての情報がすばやく手に入る情報源であることを意味している。こうしたメディアとしてのブログの特徴を生かすためには、いかに新しいトピックを取り出すことができるか、そして取り出したトピックを個人に提示して、新しいトピックについてより多く書き込んでもらえるかが重要になる。

現在では、エントリを時系列に並んだ情報として扱い、これらの出現間隔に着目することで検出する処理はよく行われている [2]。例えば、「瞬！ワード」[*5]などは、検索エンジンで検索の際に使用される検索語を、リアルタイムに集計し、急激に検索回数が増えているキーワードを掲示している。しかし、ユーザーに情報の新しさを示したり、より新しいトピックについて書き込んでもらうためには、過去の情報を示すことに加えて、ユーザー間でどのような広がり方をしているのかを分析し、用いることが重要である。本研究では、過去の統計情報とキーワードの伝播情報を用いることで、流行しているキーワードを掲示する手法を構築する。

流行しそうなキーワードを含んだエントリは、ユーザーも見たい可能性が高いと考えられるので、検索や推薦の精度の向上、ユーザー満足度の向上につながると期待できる。

連絡先: 岡田 健，東京理科大学大学院 理工学研究科 情報科学専攻，〒 278-0022 千葉県野田市山崎 2641，04-7124-1501(3329)，take@mt.is.noda.tus.ac.jp

∗1 http://www.doblog.com
∗2 http://www.cocolog-nifty.com/
∗3 http://www.blogger.com/
∗4 http://www.movabletype.org/
∗5 http://www.nifty.com/search/shun/

2. キーワード抽出アルゴリズム

ユーザーの間で流行している語、つまりユーザーが語を初めて知ったときに、語を使いたくなるようなキーワードを抽出することが目的である。そこで抽出するキーワードを

「語の出現頻度が上昇傾向かつ反応が得やすい語」

とした。つまり、ユーザーが語を閲覧する機会が増えれば増えるほど、多くのユーザーから反応を得られる機会が増える。ユーザーから反応を得られれば、よりその語が使われる機会が増える可能性が高まる。そのために、出現頻度が上昇傾向にある語を、同じ時期に多くのユーザーが語を使い始めた結果、全体として上昇傾向にあるということとした。また、ユーザーが発信したエントリを閲覧したユーザーは、コメントを書くことがある。このコメントを、発信した情報に対する反応として捉え、このコメントを異様に多く得ている語を反応が得やすい語とした。

2.1 出現頻度の上昇傾向による抽出アルゴリズム

出現頻度が上昇傾向にある語を抽出するためには、相対的にどの程度使われるようになったかが重要である。そのために、語の出現頻度が上昇傾向を、語の順位[*6]が上がっている度合い

∗6 語を出現頻度で降順に並べたときの順位

とした。その度合いに出現頻度の大きさの重みを加えた値が高い語を抽出する。このことから語の出現頻度を期間ごとに数えるとき、語 w の出現頻度を $freq(w)$ とする。現期間での出現頻度による順位を r^{now} とする。一つ前の期間での順位を r^{prev} とするとき、語 w の上昇幅は

$$up(w) = \frac{r^{prev}}{r^{now}} * (log(freq(w)) + C)$$

と表すことにする。r^{prev} を r^{now} で割った値は、出現頻度の順位という客観的な物差しを用いることで、ユーザー間で急激に使われる度合いの急激さを定量化することを意味する。一方で、出現頻度が低い語は、同順位が沢山存在するので、出現頻度が少しの変動でも非常に変動したかのような問題がある。そのために、出現頻度の対数をとった値と定数 C を重みとして加える。出現頻度は、順位の上昇率より数値的に大きい傾向にあるので、重みを小さくするために対数をとった値とした。

2.2 反応が得やすさによる抽出アルゴリズム

反応が得やすい語を抽出するには、語の出現間隔の他に、語に対するユーザーからの反応も考慮することが必要ある。そこで語がエントリに出現することによって、異常にコメントされるような語を抽出すれば良い。

語が出現することでユーザーからコメントを異様に得られる度合いを測る手法として χ^2 値を用いる。そこで、語が出現するエントリにコメントがあるエントリ数を観測値 O_i とする。語に関係なくエントリにコメントがある確率から、語がエントリに出現することで考えられるエントリ数を理論値 E_i とする。

$$\chi^2(w) = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

この観測値と理論値を用いて上記の式から求められた値が χ^2 値で、これらをエントリに語の出現の有無、そしてエントリへのコメントの有無の 4 つの組み合わせを全て調べる。この値が高い語を異常にコメントされる語として抽出する。

2.3 頻度と反応を考慮した抽出アルゴリズム

ユーザー間で流行している語を抽出するには、頻度という指標を用いることに加えて、その語がどれほどユーザーから注目されているかという視点が必要となる。このことから頻度のアルゴリズムと反応のアルゴリズムを併用したアルゴリズムである頻度と反応を考慮した抽出アルゴリズムを考える。そこで語が上昇した上昇率にコメントの有無による重みをつけることを行う。つまり、頻度と引用を考慮した抽出アルゴリズムで計算した値は

$$score(w) = up(w) * (log(\chi^2(w)) + C)$$

とした。語の上昇率が高くユーザーが閲覧する機会が増えたとしても、ユーザーが語に興味を持たなければ語は流行することはない。そのことから $up(w)$ と $log(\chi^2(w)) + C$ を乗算した値とした。

3. 評価実験と考察

前章までに提案した手法の有効性を評価実験を通じて確かめる。2004 年 4 月 1 日から 4 月 30 日までの 1ヶ月間に流行った語の抽出精度について評価実験を行う。抽出対象となるデータは、Doblog に 2004 年 4 月 1 日からの 1ヶ月間に投稿されたデータを対象とする。そして、2004 月 4 月 1 日からの 1ヶ月を計算期間として、2004 月 4 月中に出現した固有名詞の中から約 60 個の固有名詞を選び、これらの語を実験対象の語とした。そして出現頻度の上昇傾向による抽出手法で抽出した語が、実際のニュースやイベントなどとの関連があることを述べる。

表 1: $Score(w)$ とコメントがある（ない）エントリ数

単語	$Score(w)$	コメントがある	コメントがない
三菱	195.341	23	102
坂口	24.384	3	25
三菱自動車	22.822	5	26
朝鮮	18.555	5	26
玉田	17.232	15	56
ひろみ	15.629	2	13
御前崎	13.292	1	11
道路公団	12.678	2	13
てるみ	12.631	84	151
佐藤琢磨	12.496	3	21
⋮	⋮	⋮	⋮
日本	4.745	1529	2982
⋮	⋮	⋮	⋮

“ 三菱 ”を含む新聞記事は「三菱ふそうトラック・バスのリコール問題で宇佐美会長の辞任」があり、“ 坂口 ”は「日本歯科医師連盟をめぐる汚職事件」があり、“ 朝鮮 ”は「北朝鮮の列車爆発事故」があり、“ 玉田 ”は「サッカーの親善試合で初ゴール」があり、“ 道路公団 ”は「道路民営化法案」があり、“ 佐藤琢磨 ”は「バーレーン GP で 5 位」などがあった。“ 三菱 ”や“ 坂口 ”など、その時点で世間的に注目を集めている語を抽出できていることが分かる。また“ てるみ ”は、1999 年頃から 2004 年頃まで UG 系の掲示板で人気のあったネットアイドルであり、関心が持続的に維持されているような語についても抽出できていることが分かる。一方で“ ひろみ ”は、芸名やニックネームで用いられているが、出現するエントリによって対象とする人物が異なるため、語の持つ意味を識別していない点が問題である。

4. まとめ

本稿では、語が得られているユーザーからの反応に着目することで、流行しているキーワードを抽出を行った。提案手法によって得られた結果は、キーワードが出現するエントリ数だけでなく、そのエントリに対するコメントが異様に存在するかが重要な点であることが分かった。一方で、語の使われている意図を汲み取ることや、定量的な評価手法による評価などが課題である。

参考文献

[1] Nielsen//NetRatings, 2004 年 11 月の月間インターネット利用動向調査結果, 2004.

[2] 藤木 稔明, 南野 朋之, 鈴木 泰裕, 奥村 学,”document stream における burst の発見”, 情報処理学会研究報告, 2004-NL-160, pp.85-92.