blog と web 日記における内容の違いに関する考察

藤木 稔明　　南野 朋之　　奥村 学
東京工業大学大学院 総合理工学研究科　　東京工業大学 精密工学研究所
{fujiki,nanno}@lr.pi.titech.ac.jp　　oku@pi.titech.ac.jp

1　はじめに

近年、blog(weblog) と呼ばれる web ページが急激に増加している。blog とは「インターネット上に個人が公開する日記の一種で、プライベートな内容を記すことよりも、むしろ話題となっているインターネット上のニュースを引用し、コメントや批評を加えたり、新しい視点を提供するといった方向性を持っているコンテンツの総称」[1] であり、blog ツールと呼ばれる blog を簡単に書くことのできるツールがアメリカで開発された事がきっかけで流行したと考えられている。アメリカで流行した直後から日本国内へも流入してきたと考えられているが、特に 2003 年に blog ホスティングサービスのココログ[2]が開始されて以降、国内の blog ページは急速に増加しており、現在は 60 万サイト以上の blog があるといわれている[3]。

一方、blog が流行する以前から国内では web 日記と呼ばれるコンテンツを持つサイトが存在していた。web 日記とは、その名の通り web 上にアップロードされた個人の日記であり、ツールを用いず直接 HTML を編集することで書かれていたり、ツールを用いていても blog ツールのような TrackBack、RSS feed、ping といった仕組みを用意していないツールで書かれているものを指すことが多い。

blog と web 日記については、異なったものであり区別すべきなのか、それとも呼称が異なるだけで同じ物であり区別すべきではないのかという点に関して blog が日本に紹介された頃には活発な議論があった。blog/web 日記を区別すべきかどうかという点における議論に結論を出すのは難しいが、それらの間に差異があるのかどうかという点に関しては実際にいくつかの研究が発表されている。例えば、山下らは「はてなダイアリー」ユーザーに対して質問を行い、その回答の分析を行っている [2]。この研究で山下らはブログ度という指標を定義し、回答を元にしてそのユーザーのブログ度を計算している。その結果、ブログ度の高いユーザーと低いユーザーではその興味の方向が異なることを明らかにした。

ところで、我々は blog の自動収集とマイニングを行うシステムである blogWatcher[1] を開発している。このシステムでは、web ページの構造解析を行うことにより、blog/web 日記の区別をすることなく収集を行っており、それらをマイニング対象として蓄積している。しかし、blog/web 日記に差異がある場合、それらを同列に扱ってマイニング対象とすることが本当に適切と言えるのかどうかという点に疑問が生じる。

そこで本研究では、いわゆる blog ツールを使って作成されたページを (狭義の)blog と定義し、blog と web 日記の間には内容的な違いがあるのかどうかを分析することを目的とする。一般に内容の違いというと様々な違いが考えられるが、特に本研究では、blog と web 日記から注目されている話題を抽出し、その分析を行うことによって内容の違いがあるのかどうかの調査を行う。

2　分析手法

blog や web 日記に対して機械的な内容分析を試みることは従来あまり行われていない。そのための素朴な方法としては、blog と web 日記それぞれの文書集合から得られる単語分布の距離を測る方法などが考えられるが、この方法では有意な結果が得られるとは考えにくい。

そこで本研究では、blog と web 日記においてどのような話題が注目されているのかの比較を行う。注目されている話題が異なるというのは、興味を持つ話題が異なるということを意味するため、blog と web 日記では書き手/読み手の性質が異なると結論づけることが可能となる。

注目されている話題を検出するための手法には我々が以前提案した手法 [3] を用いる。この手法は、blog(web 日記) などを日付の付いた文書列 (document stream) と考え、その文書列中で他と比べて文書間間隔が短くなっている部分を検出する手法である。この手法を blog に適用した実験では、高精度で注目されている話題を表す単語 (以降ホットキーワードと表記する) を抽出可能であることが確認されている。例として、この手法を用いて 2004 年 4 月のホットキーワードを抽出した結果を表 1(データセット all: 後述) に示す。

[1] http://internet.watch.impress.co.jp/www/article/2002/0725/salon.htm
[2] http://www.cocolog-nifty.com/
[3] http://ping.bloggers.jp/ の登録サイト数より

表 1: ホットキーワードの例 (データセット all:2004 年 4 月)

1	人質	6	みどりの日
2	エイプリルフール	7	アルジャジーラ
3	邦人	8	高遠
4	光輝	9	花見
5	ゴールデンウィーク	10	植草

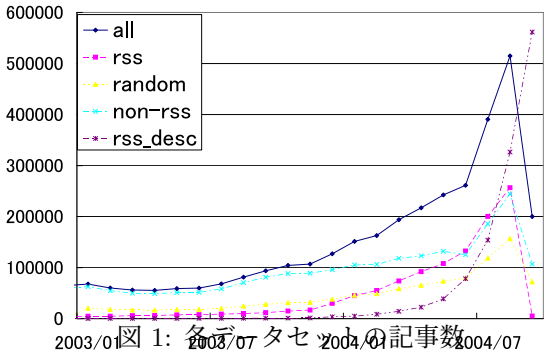

図 1: 各データセットの記事数

3 用いるデータセット

分析対象となる blog/web 日記を収集する際、blog と web 日記をどのように定義、区別するかという点が問題となる。そこで本研究では狭義の blog という分類を考え、RSS を出力するツールで書かれたページを blog と考え、そうでない物を web 日記と考える。この分類方法では、表層的な分類に過ぎないため本来の意味での blog と web 日記が分類されるわけではないと考えることもできるが、性質面での違いが実際に存在するかどうかはわからないため、このような分類方法を用いている。

具体的な分類方法としては、我々の開発した blog/web 日記収集システムである blogWatcher が収集した blog/web 日記ページのうち、RSS 配信を行っていることが宣言されているページ (RSS auto discovery [4] に対応しているページ) を blog であると考え、それ以外のページを web 日記であると考えることとする。

本研究では分析対象として以下の 5 種類の性質の blog/web 日記を、2003 年 1 月〜2004 年 9 月の期間に関して用意した。以降の分析では、これらのデータセットから計算されたホットキーワードの中で、2004 年 4 月〜8 月の物を使用する。

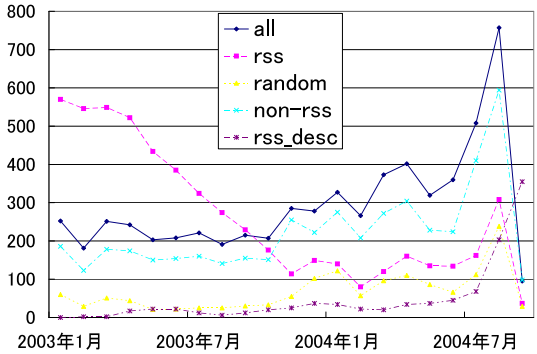

図 2: 各データセットから得られるホットキーワード数

[4] http://diveintomark.org/archives/2002/05/30/rss_autodiscovery

1. all
   blog/web 日記の区別をせず、混在させたままの文書集合。non-rss と rss の和集合に等しい。
2. non-rss
   上述の定義により web 日記とされるページを集めた文書集合。
3. rss
   同様に上述の定義で blog とされた文書の集合。rss-desc とは異なり、blog 記事のテキスト全体を用いている。
4. random
   all の中からランダムに文書を選び、総文書数が rss と同程度になるように調節したデータセット。blog と web 日記が混在。
5. rss-desc
   上記の 4 種のデータとは別に RSS を収集し、その description 要素のみを利用して構成したデータセット。収集時期・方法が異なるので、rss とは含まれる blog ページが一致していない。

   description 要素とは、各 blog 記事の要約として RSS 中に含まれているテキストである。多くの場合は記事の先頭数バイトのみのテキストであり、記事全文を含むことは少ないが、多くの blog 検索エンジンではこの部分のみを検索対象として indexing している。

これらのデータセットが含む blog 記事数を図 1 に示す。図よりわかるように、データセット non-rss の記事数はそれほど変化がないのに対し、データセット rss の記事数は増加し続けている。これは、RSS を出力する blog ツールの普及に影響を受けていると考えられる。また、RSS に含まれる情報は最近更新のあった記事だけであるため、RSS のみを収集している rss-desc は過去の記事を取得することが不可能となる。そのため rss-desc には古い記事がほとんど存在しない。

また、これらのデータセットから計算されるホットキーワード数を図 2 に示す。データセット rss において 2003 年前半のホットキーワードが多数出現しているのは、収集されたデータの中にその時期の物が少なく、同じような内容の物が複数収集されるなど、データが偏っていることの影響を受けているためである。

4 内容分析

本研究では以下に述べる 5 つの点について分析を行う。各分析において必要となる主観的判断については、著者の 1 人が行った。

4.1 ホットキーワードの精度評価 (ゆるい判定基準の場合)

まず最初に、各データセットから計算されたホットキーワードの精度評価を行う。精度評価は、2004 年 4〜8 月に関して得られたホットキーワードのそれぞれ上位 30 語を評価し、その月に話題となった単語として適当かどうかを判断することによって行われる。このようにして評価を行った場合に高い精度が得られるデータセットは、実際に注目されていた話題を多く含むということができ、正しく流行を反映しているということができる。これはまた、そのデータセットにおいてデータ収集上の偏りが少ないということも意味し、ホットキーワードの計算対象として適当であると考えることができる。

ただし各ホットキーワードについて、その月のホットキーワードとして適当かどうかの判断を与えることは難しい。そこでここでは、他の月に起こった出来事に関連する単語、収集上の偏りに影響されていることが明白な単語などのようにホットキーワードとしては明らかに誤りである単語のみを不正解とした。

その結果は表 2 のようになった。

表 2: ホットキーワードの精度 (ゆるい判定基準の場合)

	all	non-rss	rss	random	rss-desc
5 月	100%	97%	97%	100%	93%
6 月	97%	93%	90%	90%	87%
7 月	100%	97%	93%	100%	97%
8 月	100%	97%	100	100%	87%
平均	99%	96%	95%	98%	91%

この表よりわかるように、明白な誤りのみを不正解とするような判定基準を用いた場合、ほとんどのデータセットで高い精度が得られた。つまり、どのデータセットにおいてもデータ収集上の偏りなどは少なく、得られた結果にはある程度の妥当性があるということがわかる。ただしこれらの中でも rss-desc のみは他のデータセットよりも有意に精度が低かった (有意水準 5%)。

4.2 ホットキーワードの精度評価 (厳しい判定基準の場合)

前節での結果では、どのデータセットでも 100%に近い精度が得られた。しかしこの判定基準では、それほど有名ではないと評価者が感じるスポーツ選手名など、評価者が知らないだけで本当に注目されていた単語なのか、そうではないのかがわからないために正解とされている単語が多い。

表 3: ホットキーワードの精度 (厳しい判定基準の場合)

	all	non-rss	rss	random	rss-desc
5 月	60%	67%	53%	63%	50%
6 月	57%	57%	37%	47%	43%
7 月	80%	87%	70%	80%	90%
8 月	80%	77%	77%	77%	63%
平均	69%	72%	59%	67%	62%

そこで、より判定基準を厳しくして評価を再度行った場合の結果が表 3 である。この表の結果を見ると、rss の結果が悪いことがわかる。特に、最も良い結果が得られる non-rss と比較した場合には、有意に悪いことがわかった (有意水準 5%)。ただし当然ながら、この判定基準を用いた場合は前節の判定基準よりも判定者の主観に影響を受ける部分が大きく、判定のゆれがある可能性がある。

4.3 ホットキーワードにいくつの話題が含まれるか

各データセットから計算されるホットキーワードがいくつのトピックを含んでいるかの調査を行う。そのために、人手で各ホットキーワードに対し 1 つずつトピックを割り当て、各月のホットキーワードではいくつのトピックを含むかを調べる。このようにしてトピック数を調べることで、各データセットがどの程度の話題をカバーしているか、また、話題に対する注目がどの程度分散しているのかを調べることができる。

同一トピックとする判定基準は、単語が完全一致する場合の他に、「入梅」と「梅雨入り」のように同義語である場合、「台風+一過」「皆既+月食」のように同一トピックを指すことが推測できる場合、一語にすべきところを形態素解析の失敗により分割されている場合を同一トピックと判断した。ただし例外として、競走馬の名前は全て同一トピックとみなした。

この判定基準を用いてホットキーワードにいくつの話題が含まれるかを調べた結果は表 4 のようになった。

表 4: ホットキーワードにいくつの話題が含まれるか

	all	non-rss	rss	random	rss-desc
5 月	15	16	14	14	12
6 月	14	16	10	14	12
7 月	14	14	12	13	12
8 月	14	14	13	13	17
平均	14.25	15.00	12.25	13.50	13.25

表中の数字は含んでいたトピック数を示している。この表より non-rss は多くのトピックを含んでいることがわかる。これは non-rss には様々な種類のページが含まれ、他のデータセットと比べて話題が発散している事を示すと考えることができる。しかし、他のデータセットと比べても統計的に有意な差はなかった。

4.4 ホットキーワードのカテゴリー

前節では各ホットキーワードに対して 1 つのトピックを割り当てた。次に、それらのトピックをカテゴリー毎に分類する。これによってデータセット毎にどのようなカテゴリーの話題が含まれやすいのかを調べることができる。

トピックを分類するカテゴリーとしては、「スポーツ」「アニメ・ゲーム・PC」「政治・経済・社会・国際」「芸能・文化」「天候」「祝日・季節行事」「科学」の 7 カテゴリーを用いる。

このカテゴリーに対するデータセット毎のホットキーワードの分布は表 5 のようになる。

表 5: ホットキーワードのカテゴリー

	all	non-rss	rss	random	rss-desc	計
スポーツ	53	31	64	50	46	94
アニメ	4	8	1	4	1	9
政治	16	20	23	18	11	32
芸能	12	11	11	9	5	22
天候	11	11	9	10	8	16
祝日	21	24	34	23	21	33
科学	0	0	1	2	0	2

表中の数字はそのカテゴリーに含まれるホットキーワードの個数である。またここでは、2004 年 4 月〜8 月のホットキーワードを全て合計した値となっている。この表の中で特に non-rss と rss に着目すると、rss では「スポーツ」が多く、non-rss では「アニメ・ゲーム・PC」が多い事がわかる。

4.5 ホットキーワードが表す話題の一致割合

最後に、各データセット間でのトピックの一致割合を調べる。そのために 4.3 節で各ホットキーワードに対して割り当てたトピックを利用し、あるデータセットに含まれるトピックが他のデータセットにどの程度含まれるかを調査する。

このトピック一致割合が高ければ、それらのデータセットの内容は似ているということができる。

結果は表 6 のようになった。この表は、行が示すデータセットに含まれるトピックが、列の示すデータセットのトピックに何%含まれているかを示している。つまり例えば、表の一番右上の数値は、all に含まれるトピックのうち 62%が rss-desc の示すデータセットに含まれていることを示している。

表 6: ホットキーワードが表す話題の一致割合

	all	non-rss	rss	random	rss-desc
all	-	0.88	0.73	0.82	0.62
non-rss	0.83	-	0.59	0.75	0.60
rss	0.71	0.63	-	0.77	0.59
random	0.79	0.77	0.76	-	0.59
rss-desc	0.58	0.59	0.59	0.59	-

この表よりわかるように、rss-desc は他のどのデータセットとも一致割合が低く、non-rss と rss の一致割合も低いという結果が得られた。

5 考察とまとめ

以上の実験からは以下のことが推測できる。

まず、どのデータセットからも高精度のホットキーワードが得られたことから、どのデータセットでもデータ収集の偏りは少なく、ホットキーワードの計算対象として適当であると考えることができる。次に、上位 30 語のホットキーワードが示すトピック数を比較すると、non-rss のトピック数が多いという結果が得られた。これは、blog よりも web 日記の方が話題が分散している可能性があることを示している。ただしこれには有意差はなかった。さらに、blog/web 日記で注目された話題はどのようなカテゴリーの話題であったかを調査すると、blog ではスポーツに関する話題、web 日記ではアニメ・ゲームに関する話題が多い傾向があることがわかった。また、blog と web 日記ではトピックの一致割合が低かった。この結果からも、blog と web 日記では同じ話題が注目されているのではないということがわかる。つまり、blog と web 日記では書かれやすい話題やそのカテゴリーが異なるため、blog と web 日記には内容の違いがあるという結論が得られた。

マイニングのためのデータという観点から本実験の結果を考えると、4.1 節の結果が示すように、rss-desc から計算されるホットキーワードの精度は有意に悪かった。これは RSS の description 要素に含まれるテキストの長さが短いことに起因していると推測することができる。よって、マイニングという目的のためには、RSS のみから得られるデータを用いるよりも、blog/web 日記の記事に含まれる全テキストを利用した方がより良い結果が得られるということができる。

しかし一方で、本分析では評価対象のデータ数が少なく、また評価者も一人しか確保できなかった。本分析は主観的な評価を行う場面も多く、評価の揺れも起きやすいタスクである。評価を行う人数を増やすなどの方法によって、より正確な結果を得ることも可能であるかもしれないが、それは今後の課題として残されている。

また一般に blog と web 日記の違いを分析しようと考える場合、本研究のようにどのような話題が注目されているかという点を分析する以外にも考慮すべき点がいくつか考えられる。例えば、文書中に出現するリンクの数や他の blog/web 日記からの被リンク数などといったリンクに関する違い、平均文書長や出現しやすい文末表現の違いなどといった文章のスタイルに関する点などがあげられる。さらに、以前は web 日記スタイルで書いていたが、その後 blog ツールを使うようになったという層に関してその記述内容の変化を調べることも有意義であると考えられる。このような点に関する違いの考察は今後の課題である。

参考文献

[1] 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 セマンティックウェブとオントロジー研究会, SIG-SWO-A401-01, 2004.

[2] 山下清美, 三浦麻子. 人はなぜウェブ日記・ウェブログを書き続けるのか (2). 日本社会心理学会第 45 回大会, 2004.

[3] 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, 2004.