人工知能学会研究会資料
SIG-SW&ONT-A401-01

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

奥村 学[†]　南野 朋之[‡]　藤木 稔明[‡]　鈴木 泰裕[‡]

概　要

blog を掲示板と同様の情報源として，定期的に監視し，そこから情報を抽出，発掘するためのシステムを開発している．ホットキーワード抽出でホットな話題をチェックする，評価表現抽出を利用した評判情報検索をする，また，お勧め blog を提案する，などの機能を特徴としている．

Automatically Collecting, Monitoring and Mining Japanese Weblogs

Manabu OKUMURA[†]　Tomoyuki NANNO[‡]
Toshiaki FUJIKI[‡]　Yasuhiro SUZUKI[‡]

Abstract

We present a system that tries to automatically collect and monitor Japanese blog collections that include not only ones made with blog softwares but also ones written as normal web pages. Our approach is based on extraction of date expressions and analysis of HTML documents. Our system also extracts and mines useful information from the collected blog pages.

1　はじめに

インターネットの普及に伴い，一般の多くの人々からの情報発信が盛んになり，その発信されている大量の情報を有効に活用したいという要求も高まっている．こうした状況を背景に，現在注目されている情報源の一つが掲示板 (BBS) であり，掲示板を定期的に監視し，そこから情報を抽出，発掘することで，一般大衆の「生の声」を製品開発，企業活動に反映しようという試みも見られる [1].

同様に近年注目され始めている情報源として blog (Weblog) がある．blog の定義は現在必ずしも定まっているとは言えないが，Web 上の「日記サイト」あるいは「個人ニュースサイト」と言うことができ，書き手が関心を持ったニュースやできごとについて (何らかのコメントを) 書いた記事を，元情報へのリンクとともに時系列に沿って掲載しているサイトを指すことが多い．通常の Web ページとは異なり，速報性，リアルタイム性のある新鮮な情報が発信されることから，掲示板同様有用な情報源と考えられるようになってきている．

現在,「blog」というと，Movable Type[1]などの「blog ツール」や，tDiary[2]などの「Web 日記システム」，また，これらのホスティングサービスを利用して書かれ，Trackback[2] など，blog ツール特有の機能を有しているサイトを指すことが多い．

これらのツールを使用した blog の収集は，RSS(RDF Site Summary / Rich Site Summary / Really Simple Syndication)[3] などのメタデータを利用したり，さらには，`ping.bloggers.jp`[3]などをはじめとする，XML-RPC を使用した ping[4] による blog 更新通知サービスを利用することで，比較的容易に行うことが可能である．

しかしながら，日本ではこうした blog ツール登場以前から「Web 日記」あるいは「テキストサイト」という形で，個人による情報発信が行われており，非常に大きなコミュニティを形成している [5]．これらの情報

[†]東京工業大学精密工学研究所
Precision and Intelligence Laboratory,
Tokyo Institute of Technology
〒 226-8503 横浜市緑区長津田町 4259 R2-720
oku@pi.titech.ac.jp
[‡]東京工業大学大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology
〒 226-8503 横浜市緑区長津田町 4259 R2-728
{nanno,fujiki,yasu}@lr.pi.titech.ac.jp

[1]movabletype.org `http://www.movabletype.org/`
[2]tDiary.org `http://www.tdiary.org/`
[3]ping.bloggers.jp　`http://ping.bloggers.jp/`

も blog と同様，注目すべき情報であるが，その多くがツールやホスティングサービスを利用しない，個人が管理する Web ページ中に存在するため，網羅的に収集することはそれほど容易ではない．

そこで本研究では，このような Web 日記も含めた，時系列に沿ってなんらかの記述を掲載しているサイトを blog と定義する．そして，特定のツールやメタデータに依存しない，HTML 文書の解析に基づいた手法で，これら個人の発信する時系列に沿って掲載された情報を網羅的に収集, 監視する．また，我々は，収集した blog から有用な情報を抽出，発掘するためのシステム “blogWatcher” を開発している．本稿では，“blogWatcher” の機能を中心に紹介する．

2 関連研究

blog を収集し，検索機能を提供しているサービスはいくつか存在する．

Bulkfeeds[4]は日本国内で RSS 配信されている情報を検索することのできる検索エンジンである．RSS の収集は以下の三種類の方法で行われている．

- ping サーバの更新情報
- blog ホスティングサービスの更新情報
- 人手による登録

また，blog のみに的を絞った検索エンジンとして，Feedback[5]や Myblog japan[6]，ココログ[7]，BLOGNAVI[8]，blog search![9]などがある．これらは，人手による収集，あるいは ping サーバから blog サイトの情報を得て，RSS を利用することで，blog を収集する検索システムである．

これに対し，本研究では，クローリングした HTML 文書を解析し，その Web ページが blog であるかを判定することによって blog の収集を行う．この手法の利点は，以下の 2 点である．

- blog ツールなど，特定のシステムを利用していない Web 日記なども広く収集することが可能
- RSS は，直前何回かの更新に対するメタデータであることが多いため，RSS ベースの収集では過去の記事を収集できないのに対し，HTML 文書に基づいた収集では，HTML 文書を直接解析することで，過去のものまで収集することが可能

[4]Bulkfeeds `http://bulkfeeds.net/`
[5]Feedback `http://naoya.dyndns.org/feedback/`
[6]Myblog japan `http://www.myblog.jp/`
[7]ココログ `http://web.or.tv/`
[8]BLOGNAVI `http://www.blognavi.com/`
[9]blog search! `http://blog.threetree.jp/`

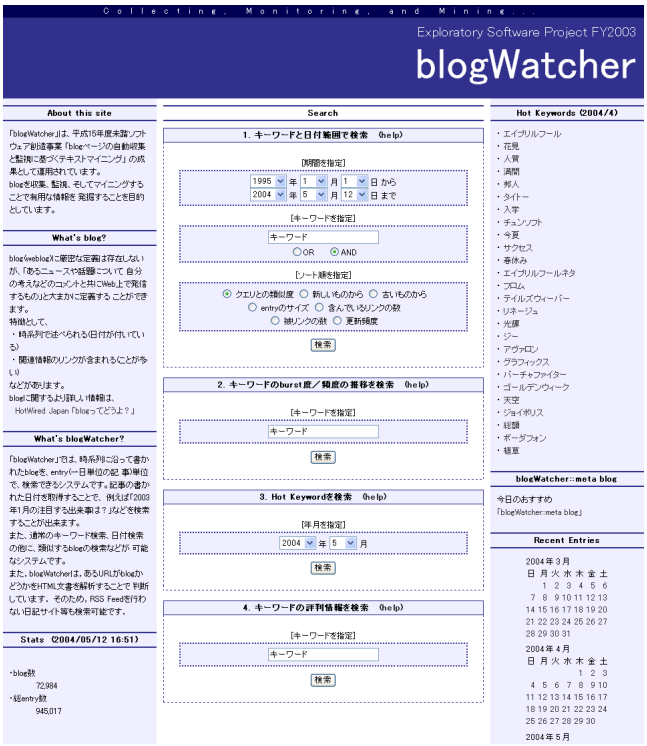

図 1: blogWatcher のスクリーンショット

また，海外では，Technorati[10]や Blogdex[11] などが RSS に基づいた検索サービスや，含まれるリンクを利用したニュースソースのランキングなどを行っている．

3 blogWatcher

“blogWatcher” は，blog を収集・監視し，また収集した blog をマイニングすることで得られた情報を閲覧することが出来るシステムである．図 1 に，システムのスクリーンショットを示す．

blog の収集・監視については，ping サーバや RSS などのメタデータに依存しない，HTML 文書の解析に基づいた手法で行っている．収集システムのフローチャートを図 2 に示す．(収集・監視システムの詳細については，[6, 7] を参照．)

現在のシステムの機能について以下に示す．

- 全文検索
  - キーワード，日付による検索
  - 被リンク数，更新間隔，サイズ等によるランキング
  - 検索結果の RSS Feed
- キーワードの月間出現数推移

[10]Technorati `http://www.technorati.com`
[11]Blogdex `http://blogdex.media.mit.edu`

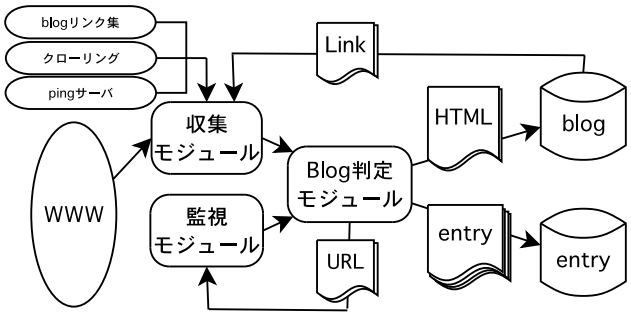

図 2: 収集・監視システムのフローチャート

- ホットキーワード抽出 (burst 検出)
- 評価表現 (主観表現) の抽出，表示
- 評判情報検索
- おすすめ blog の提案

次節以降で，主要な機能について詳細に説明する．

3.1 blog 検索

blogWatcher では，収集モジュールが収集した HTML 文書に対して blog 判定を行い，blog と判定されたページについては，日付表現を手がかりに一日分の記事単位に分割する．そのため，blogWatcher では，この分割された一日分の記事 (以降，entry と呼ぶ) を基本単位として検索可能である．

なお，検索エンジンは，情報処理振興事業協会 (IPA) が実施した「独創的情報技術育成事業」の研究成果である “汎用連想計算エンジン (GETA)” [8] を使用した．GETA を採用した理由は，以下の二点の特徴が blog 検索を行う上でユーザにとって非常に有用であると思われたためである．

- 検索結果に特徴的に現れる単語 (トピックワード) の抽出が容易 (図 3)
  (トピックワードを用いることで，ユーザは大量の entry の中から自分の欲しい情報を容易に絞り込み検索することが可能になる．)
- 文書をクエリとした関連文書検索が可能
  (ユーザは自分の興味のある entry を選択するだけで，類似する entry を検索することができる．)

以下では検索機能について説明する．

全文検索・日付検索

ユーザは，以下の二種類の方法で entry を検索することが可能である．

1. 検索キーワードを指定する
2. 日付の範囲を指定する

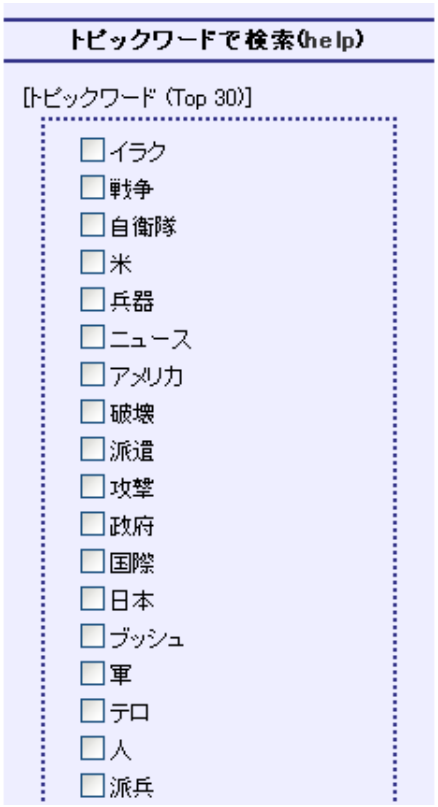

図 3: 「イラク」で検索した際のトピックワード

また，これらを組み合わせて使用することもできる．さらに前述したように，検索エンジンが提示するトピックワードを使用した “トピックワード検索”，興味のある entry に類似する entry を検索する “関連文書検索” も可能である．

なお，entry の日付は，Web ページ中に記述される日付表現を自動的に抽出することにより得ている．日付表現の区切り文字には，年月日，ハイフン，スラッシュなどが使用されることもあれば，月が英語名で記述されていたり,「平成」のような年号をつけて記述されていることもある．このような日付表現のフォーマットの多様性に対応するために，人手で記述した正規表現を使用している．また，blog では，最上部に年が記載され，各 entry の日付では年を省略し，月日のみを記載することがある．このような場合でも，ヒューリスティックに基づき，不完全な日付表現に対して，不足している情報を補完することを試み，完全な年月日を取得している．

ランキング

現在実装されているランキング手法は，以下の 7 通りである．

1. クエリとの類似度

tf*idf に基づく類似度計算をすることにより，クエリとの類似度順に entry を提示.
(よりクエリに関連のあるものから順に見たい．)

2. 新しいものを重視
日付が現在に近い方から順番に entry を提示.
(最近の entry から順に見たい．)
3. 古いものを重視
日付が現在から遠い方から順番に entry を提示.
(時系列順に見たい．)
4. entry のサイズ
文字数の多い順に entry を提示.
(情報量の多い順に見たい．)
5. 含んでいるリンクの数
含んでいるリンクの多い順に提示.
(より多くのポインタを示している entry から見たい．)
6. 被リンクの数
被リンクの多い順に entry を提示.
(より多くの blog からリンクを受けている entry から順に見たい．)
7. 更新頻度
更新頻度が高い blog に含まれる entry から順に entry を提示.
(より頻繁に更新される blog の entry から順に見たい．)

検索結果の表示

図 4 に検索結果の表示例を示す．指定された検索条件 (キーワード，日付) にマッチする entry が，指定されたランキング方法に従って表示される．個々の entry に関する情報として，以下の情報が表示される．

- entry を含む Web ページの URL，タイトル
- 認識された日付情報
- 指定されたランキング方法に基づき計算されるスコア
- entry に含まれる肯定的/否定的な評価表現の数 (3.4 節を参照)

また，検索結果の一覧では，以下の情報へのアクセスが可能である．

- Zero-Click[9] を使用した，entry のプレビュー
- リンク解析結果 (他の entry へのリンクやその entry へのリンク一覧)

検索結果から，entry を選択すると，以下の情報が提示される．

図 4: 検索結果の例

- entry のスナップショット (キャッシュ表示)
評価表現 (肯定/否定) 及びクエリのハイライト表示
- オリジナルのページへのリンク
- その blog に特徴的なトピックワード
- 同じ URL に含まれる別の日付の entry 一覧，及び各 entry のトピックワード

類似文書判定

現在のシステムは，URL 単位で entry を管理しているため，別 URL にほぼ同じ内容の entry が存在してもそれらを自動的に同一であると判定する事は出来ない．そこで，ユーザが指定した閾値以上類似した entry については，図 6 のようにまとめて表示する機能を別途用意している．なお，類似度は entry の単語ベクトルの余弦で計算され，閾値は 0.3〜1 の間で自由に設定が可能である．(使用しないことも可能．)

この機能により，ユーザは類似する entry を読み飛ばすことが出来るようになるため，より効率の良いブラウジングを行うことが出来るようになる．

検索結果の RSS Feed

blogWatcher では，検索結果を RSS で受け取ることが可能であり，この API を公開する予定である．(図 7) よって，RSS Reader を使用すれば，自分の興味のあるキーワードを含んだ entry を定期的にチェックすること

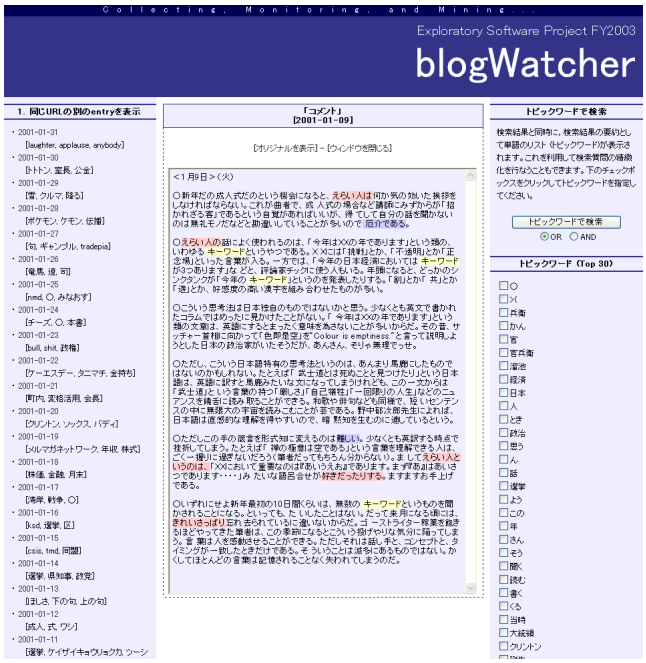

図 5: entry のスナップショット

図 6: 類似文書判定

が容易になったり，また，procfeed[12]などのサービスと組み合わせることで，検索結果を blog や Web ページに挿入することも可能である．

3.2 注目度 (burst 度) の表示

ある程度の量の blog ページ集合を利用することが可能な場合，それらのページの中で，あるキーワードの出現頻度がどのように推移するかを測るなどすることで，そのキーワードが「いつ」「どの程度」注目されていたのかを知ることが可能である．しかし出現頻度が閾値以上になっている期間を抽出するなどといった単純な方法では， キーワード毎の総出現頻度の差に影響を受けたり，閾値付近で値が変動する場合に小さな注目期間が大量に検出されることになるという問題がある．そこで本システムでは Kleinberg の提案する手法 [10] を利用して burst を発見することによって，注目されている期間を検出している．この手法では blog ページ集合の中である期間が burst 状態であるかどうかを，

[12]procfeed `http://procfeed.net/`

```
- <item rdf:about="http://homepage.mac.com/naoyuki_hashimoto/iblog/20031229135319">
  <title>Letter from Yochomachi [2004-05-26]</title>
  <link>http://homepage.mac.com/naoyuki_hashimoto/iblog/20031229135319</link>
  <description>「阪神・サッカー・ハルウララ」（堺屋太一）今週の週刊ポストで堺屋太一は書きました：（今の２０代３０代は）私は「阪神・サッカー・ハルウララ」と呼んでいるんですが、一生懸命頑張っているんだから、負けてもいいじゃないかと思っ ...</description>
  </item>
- <item rdf:about="http://ha.prosou.nu/diary/bin/diary.cgi?year=2002&mon=5">
  <title>GWA's WEB [2004-05-25]</title>
  <link>http://ha.prosou.nu/diary/bin/diary.cgi?year=2002&mon=5</link>
  <description>... 買い込み帰宅。鮭焼いて、いかの缶詰と味噌汁で飯。今日はだいぶリッチ。残りご飯を全部たいらげる。あまりに満腹でサッカー観戦途中で昼寝。起きたら 11 時。喉痛い。明日はひさびさに何もない日曜日。そういえば、昨日 ...</description>
  </item>
- <item rdf:about="http://plaza.rakuten.co.jp/tamachan2003/diary/">
  <title>楽天広場（日記・ブログ）:日記 生活を楽しむページ - 不況で気分も落ち… [2004-05-25]</title>
  <link>http://plaza.rakuten.co.jp/tamachan2003/diary/</link>
  <description>... いつもの独り言でした?それにしても、男子バレー惜しかった!あそこまで行って、悔しいでしょうね。夫と長男は、野球かサッカーしか見ないため、次男と二人で見てました?コメント (3)|トラックバック (0) コメントを書く ...</description>
  </item>
- <item rdf:about="http://www.ii-okinawa.ne.jp/people/l-rakara/">
  <title>汚れたての Angel [2004-05-22]</title>
  <link>http://www.ii-okinawa.ne.jp/people/l-rakara/</link>
  <description>... ーーーー！間違い探し。あせっちゃうって …。死ぬ前に、凄い生命力で乗り切ります！オタクでお茶目な、サッカー選手。当てちゃ、イヤっーーーーーーー！ ...</description>
  </item>
- <item rdf:about="http://www3.vis.ne.jp/~asaki/p_diary/diary.cgi?">
  <title>新・たけぞう瀕死の日記 [2004-05-16]</title>
  <link>http://www3.vis.ne.jp/~asaki/p_diary/diary.cgi?</link>
  <description>... 結局連休中にやろうと思っていたことを全て終わらせることはできませんでした。なむー。■ [サッカー] ゾルダーの降格が決定鈴木の古巣であり、ゾルダーとの関係も強いらしいゲンクとの対戦。鈴木の先制点 ...</description>
  </item>
```

図 7: 検索結果の RSS Feed の例 (一部抜粋)

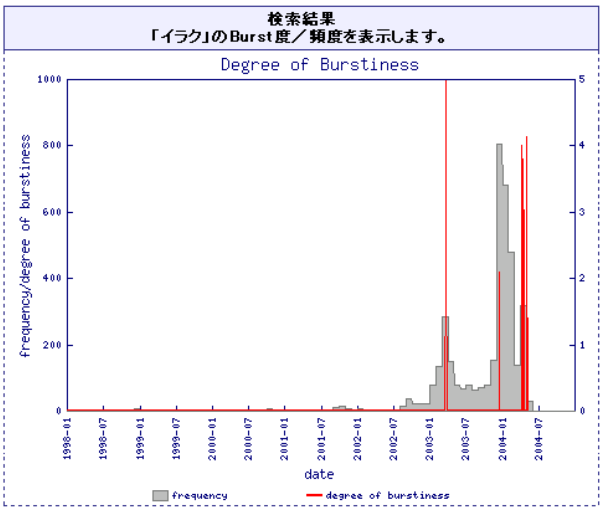

図 8: 「イラク」に対する出現頻度と burst 度のグラフ

その期間での blog の出現間隔が通常よりもどの程度短くなっているかを元として計算している．ただし blog を対象として計算する場合，収集された blog 数が時間軸に対して均一ではないため，Kleinberg の手法をそのまま適用することができないという問題がある．そのため，本システムでは我々が拡張した手法 [11] を利用して burst の発見を行っている．

blogWatcher はこのような burst の度合を，出現頻度の推移と共にグラフ化して表示する機能を持っている．図 8 は，キーワード「イラク」を入力した際表示されるグラフである．グラフの横軸は時間軸になっており，薄い色の棒グラフが出現頻度の推移を，折れ線グラフが burst 度の推移を示している．このグラフにおいては，

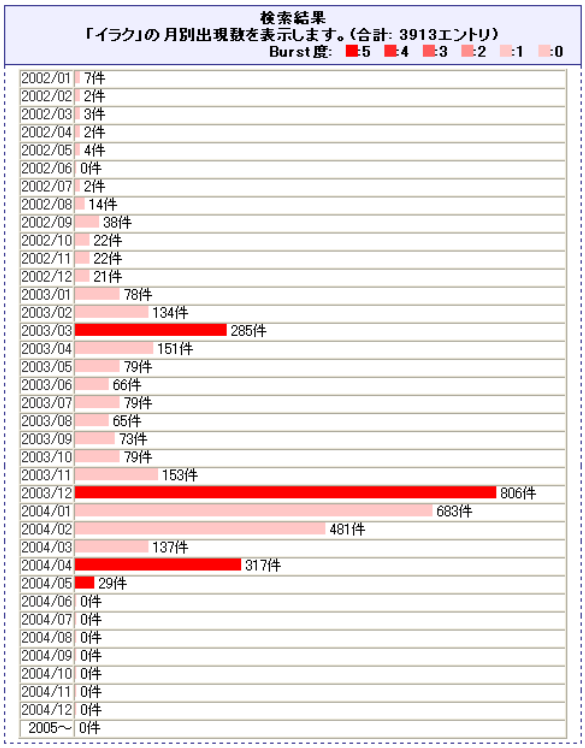

図 9: 出現頻度と burst 度の詳細表示

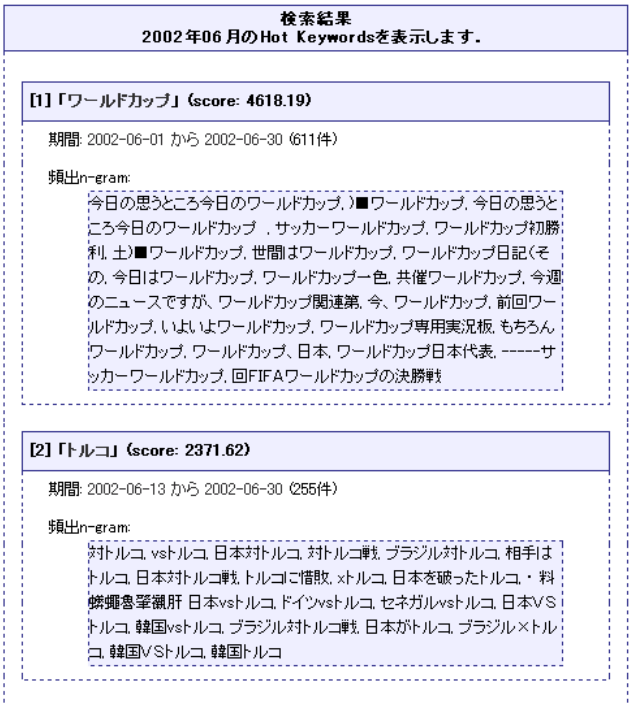

図 10: 2002 年 6 月のホットキーワードリストの一部

2003 年 3 月，2003 年 12 月，2004 年 4 月といった部分で burst が起こっていると判定されていることが折れ線グラフからわかる．実際その部分の blog を見てみると，これらはそれぞれ「イラク戦争開始」「フセイン拘束」「日本人拉致事件」に関連する blog が増えているためであることが確認できる． また，blogWatcher では，このグラフの他に，前節で説明したトピックワードの表，棒グラフ形式による出現頻度と burst 度の推移 (図 9) などが表示され，それぞれからのリンクによって，例えば,「その月のその burst 単語を含んでいる entry」など，関連する検索を容易に行うことができるようにもなっている．

3.3 ホットキーワードリストの表示

前節で述べた burst 度をあらかじめ全ての単語について計算し，それらを月毎に集計することによって，各月ではどのようなキーワードが注目されていたのかをホットキーワードリストとして表示することが可能となる．図 10 は 2002 年 6 月のホットキーワードを出力した際のスクリーンショットの一部である．図ではホットキーワードとして「ワールドカップ」「トルコ」が表示されているが，この月はサッカーのワールドカップが開かれていた月であるため，これらに続くキーワードも「チュニジア」「ベルギー」「決勝」「ブラジル」「イングランド」などのようなキーワードとなっている．また図に示されているように，ホットキーワードはそのキーワードが burst していた期間，その周辺で頻出する文字列 (頻出 n-gram) という情報と共に表示されている他，ホットキーワードを含む blog 記事を実際に検索することも可能になっている．

ホットキーワードの表示は，特定の時期にどのような話題が注目されていたのか知るための手がかりとなりうる．前節では，あるキーワードを入力として与え，出力としてそのキーワードが注目されていた時期を得る機能について述べたが，本節の機能では，時期を入力として与えることで，その時期に注目されていたキーワードを得ることができる．

ただし現状ではホットキーワードとして表示される語は 1 形態素のみとなっている．そのため「ワールドカップ」のように出来事を想起することが可能なキーワードだけでなく，キーワード単体では意味をなさない「ソル」(ソルトレイクシティオリンピックの一部) のような形態素が (形態素解析誤りにより) キーワードとして出力されることもある． そのため，キーワードに隣接する頻度の高い文字列を同時に出力することで，そのキーワードがどのような話題に関連しているのかを推測しやすくしているが，十分とは言えず，今後の改善が必要な部分であると考えている．

3.4 評判情報検索

blogWatcher では，ある対象 (キーワード) に関する評判情報を検索することができる．検索は (通常の blog 検索と同様に) キーワードを入力し,「評判情報検索」をクリックすることで行われる．検索結果は，キーワードに対する何らかの評価を行っている文が，記事毎にまとまって一覧表示される．

気になる評価文が見つかった場合，記事のタイトルをクリックすることで，その文を含む実際の記事全体を表示することもでき，より詳細な情報を得ることができる．

実際にキーワードとして「レストラン」を入力して検索を行った結果を図 11 に示す．評価表現が肯定的な評価か，否定的な評価かも自動的に判定することができ，肯定的な表現は背景が赤く，否定的な表現は背景が青くハイライトして表示されている．また，評価表現ではあるが，肯定的か否定的かの判断ができない場合は背景が緑色にハイライトして表示される．

ある文がキーワードに対する評価を行なっているかどうかを判定する際には，キーワードと評価表現を単に含む文というだけでなく，文中の係り受け関係を考慮することにより，評判情報らしさの高い文のみを抽出している．例えば,「今日は，レストランに行った帰りに，駅前のおいしいラーメン屋でも食事をした.」といったような「レストラン」というキーワードも評価表現も含まれているが，レストランに対する評判ではない文や,「このレストランが良ければ問題ないのだが … 」というように，仮定を表していて評判ではない文を除外することができる．これにより，目的とするキーワードが評価されている文だけを抽出し，提示することができ，効率よく評判情報の収集を行うことができる．

肯定的な評価であるか，否定的な評価であるかを判定するにあたっては,「おいしい」や「まずい」のように評価語だけを見れば肯定/否定が判定できるものもあるが,「あの HDD は容量が大きい」と「あの HDD は動作音が大きい」という 2 文における「大きい」が，肯定的な評価であったり否定的な評価であったりするように，何を評価しているかまで考慮して判定を行わなければならない場合がある．本システムでは，評価対象や評価の際の着目点を考慮に入れて肯定・否定の判定を行うことにより，より精度の高い肯定/否定の判定を行うことが可能となっている．

評価表現の抽出や肯定/否定の判定に使用している辞書は，現在は人手で作成した小規模なものであるが，今後，機械学習を用いて拡張 [12] を行う予定である．

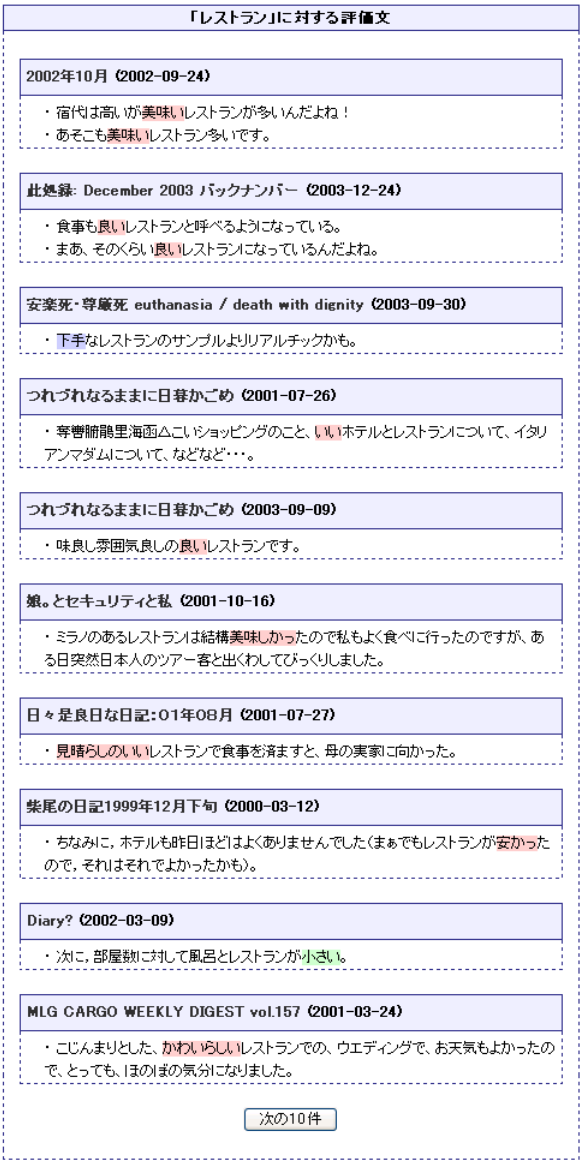

図 11: 評判情報検索のスクリーンショット (キーワード「レストラン」で検索)

3.5 blogWatcher::meta-blog

blogWatcher では，毎日おすすめの blog やホットな話題を提示するための blog を自動生成している．これは，全単語について burst を計算することでその日に burst している単語のリストが得られるため，その単語を元に注目されている話題を発見し，その話題を「blogWatcher::meta-blog」という名称の Movable Type で作られた blog に対し，毎日記事を投稿することで実現されている．

図 12 にスクリーンショットを示す．ここでは 2004 年 5 月 5 日の entry が表示されているが，その日に注目されている話題を示す語句と，その語句を含む blog の中で被リンク数の多い blog の抜粋が生成されている．

図 12: blogWatcher::meta-blog

また，それぞれの抜粋の全文を表示するためのリンクや，その語句を含む blog 記事を検索することなどができるようにもなっている．

このように meta-blog を眺めるだけで，blog 上ではどのような話題が現在注目されているのかを簡単に知ることができる．つまり，他の機能ではユーザが自分の欲しい情報を得るためにクエリを入力する必要があるのに対し，meta-blog では特に何も入力しなくても，システム側が推薦するコンテンツを見ることができるというような使い方の違いが存在する．また，meta-blog に表示される情報はホットキーワードよりも即時性が高く，短期間の話題を重要視する傾向にあるため，その日にどのような話題が注目されていたのか知りたいという目的のために利用可能となっている．

4 おわりに

blog を掲示板と同様の情報源として，定期的に監視し，そこから情報を抽出，発掘するためのシステム blogWatcher について紹介した．現在 blogWatcher の公開に向けシステムの brush-up を行なっているところであり，今夏の公開の予定である．なお，blogWatcher に関する情報は，`http://www.lr.pi.titech.ac.jp/blogwatcher/` から参照可能である．

公開後は，有用と考えられる他の機能を blogWatcher に追加し，システムを拡張していくことを検討している．

謝辞

本研究の一部は，独立行政法人情報処理推進機構 (IPA) 「未踏ソフトウェア創造事業」喜連川 優 PM による「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」の成果の一部である．また，本研究の一部は文部科学省科学研究費 (21 世紀 COE プログラム「大規模知識資源の体系化と活用基盤構築」) の補助のもとに行われた．

参考文献

[1] 松村真宏. チャンス発見のためのコミュニティマイニングに関する研究. 平成 14 年度東京大学大学院工学系研究科電子工学専攻博士論文, 2003.

[2] Benjamin and Mena Trott. Trackback technical specification. `http://www.movabletype.org/docs/mttrackback.html`, 2002.

[3] Dan Libby. Rdf site summary (rss) 0.9 official dtd. `http://my.netscape.com/publish/formats/rss-0.9.dtd`, 1999.

[4] D Winer. Weblogs.com xml-rpc interface. `http://www.xmlrpc.com/weblogsCom`, 2001.

[5] yomoyomo. Hotwired japan - 日本における blog の過去・現在・未来. `http://www.hotwired.co.jp/matrix/0305/004/index.html`, 2003.

[6] Tomoyuki NANNO, Yasuhiro SUZUKI, Toshiaki FUJIKI, and Manabu OKUMURA. Automatic collection and monitoring of japanese weblogs. In *Proc. WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.

[7] 南野朋之, 鈴木泰裕, 藤木稔明, 奥村学. blog の自動収集と監視. 情報処理学会研究報告, 2004-NL-160, pp. 129–136, 2004.

[8] 情報処理振興事業協会 (IPA). 汎用連想計算エンジン (geta). `http://geta.ex.nii.ac.jp/`, 2002.

[9] Tomoyuki NANNO, Suguru SAITO, and Manabu OKUMURA. Zero-click : a system to support web browsing. In *Proc. The Eleventh International World Wide Web Conference*, 2002.

[10] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1–25, 2002.

[11] 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, 2003-NL-160, pp. 85–92, 2004.

[12] 鈴木泰裕, 高村大也, 奥村学. Weblog を対象とした評価表現抽出. 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-A401-02, 2004.