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

1. 背景

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．こういう状況を背景に，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．

同様に近年注目され始めている情報源として blog(Web log)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．

掲示板は，その数もあまり多くなく，そのため，定期的な監視を網羅的に行うこともそれほど困難とは考えにくい．一方，blog は掲示板と異なり，サービスとして運用されている(したがって，多数の書き手が書いた記事をまとめて収集できる)ものは(特に日本では)それほど多くなく，多くは通常の Web ページと変わらず個人が各自書いているものが多数を占めている．そのため，定期的な監視を網羅的に行うことはそれほど容易ではない．

2. 目的

そこで本プロジェクトでは，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する．システムは以下の 3 つのモジュールから構成される．

1) blog ページとして監視するべき URL の特定，自動収集

blog ページの属性と考えられる情報を利用して，WWW 上をクローリングすることで得られたページ集合から，blog ページのみを選択的に自動収集する．

2) blog ページの定期的監視

収集した blog ページ集合を定期的に監視し,更新された部分のみを選択的に抽出する．

3) 内容に基づく分類，テキストマイニング

2)で収集した blog ページ(の更新部分)を，内容を元にグループ化し分類する．そして，分類した blog ページ集合ごとに，テキストマイニングを行い，有用な情報を抽出，発掘する．

3. 開発内容

システム構成は図 1 の通りである．

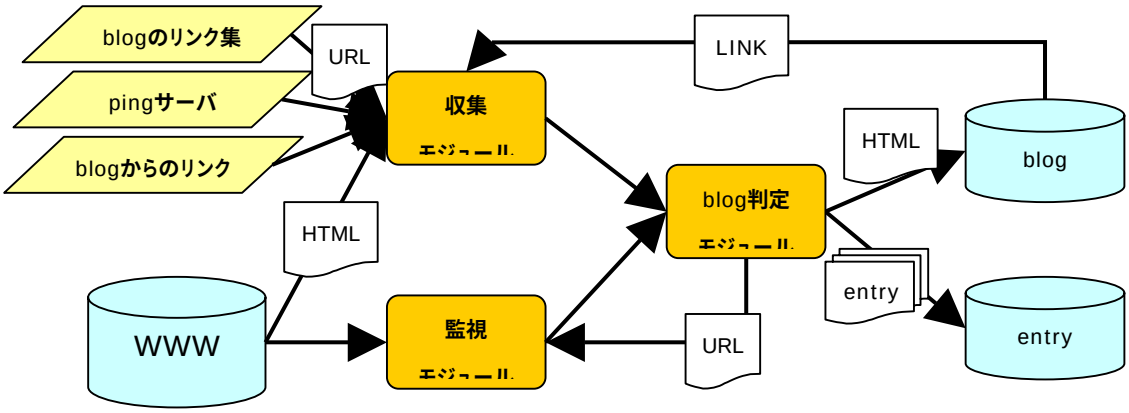


図 1:システム構成

収集モジュールは，以下の 3 つの方法で Web ページのクローリングを行う．

- WWW 全体を対象とした収集
- blog のリンク集や，更新通知サービスを利用した収集
- blog と判定されたページに含まれるリンクを利用した収集

収集モジュールが収集した Web ページに対して，blog 判定モジュールはそのページが blog であるかどうかを判定し，blog であると判定されたページから，entry を抽出する．監視モジュールは，判定モジュールによって blog と判定された URL に対して，定期的な監視を行い，更新がある度に新たに追加された entry の取得を行う．

収集した blog ページおよび entry は，キーワード，日付などで検索でき，

- entry の大きさ
- 被リンク数
- 更新頻度

などでランキングが可能である．また，リンクに関する情報が閲覧可能である．さらに，検索結果には entry の内容が類似するものが含まれてしまうケースが存在するので，類似する entry をまとめて表示する機能を実装した．

以下開発内容についてさらに述べる．また，システムを用いた検索の実行例を 3.4 節で示す．

3.1. blog ページ選択

blog 判定モジュールでは，以下の 2 つのタスクを行う．

1. Web ページ集合から blog ページを選択する
2. 選択された blog ページから entry の集合を切り出す

ここで，entry とは，blog ページに記載されている，日付によって見出しが付けられている一日分の記事のことを示す．

blog ページの選択および entry の抽出を行う際に,blog の性質として以下の性質を仮定している.

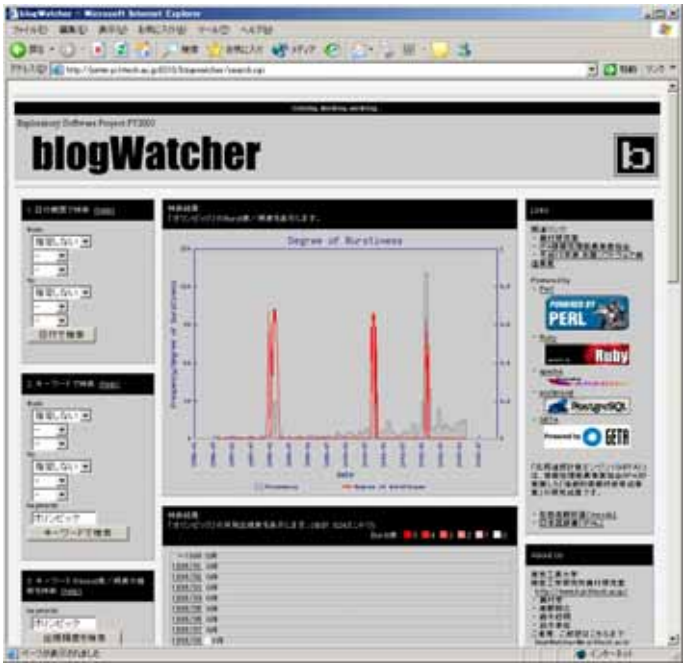
- 日付情報は必ず含まなければならない
- 日付が記事の上部にある
- 日付部分は規則正しく書かれている
  - 日付部分に関してタグの係り方は一定,
  - 日付の書き方も一定

上で述べたように,我々は日付情報を一定の規則で含むことを blog ページの性質として用いている.そこで, blog 選択で必要な日付表現を html 文書中から抽出する手法を開発した.また,blog ページの選択は,ある Web ページから entry 集合の候補となるものを抽出し,それが別途用意したフィルタリング条件に合致するかどうかで行う(合致しなければ blog と判断される).なお,blog ページ選択の詳細は,[南野,2004]を参照して欲しい.

3.2. テキストマイニング

3.2.1. キーワードに関する burst 度の計算と表示

本節では,キーワードが与えられた時に,そのキーワードの burst 度を計算,表示する機能について述べる. burst 度とは時系列に整列した文書群(document stream)から,キーワードの出現が活発になっている(盛り上がっている)個所を発見する指標である.我々は[Kleinberg,2002]が提案する手法を拡張した手法を用い,指定されたキーワードを含む blog 記事集合から burst 度の計算を行う.これによって,どの時期にそのキーワードに関する話題が盛り上がっていたのかを知ることができる.また,インタフェースからキーワードを指定して呼び出した際に,burst 度をグラフとして表示する機能を実装した(右図).



この図ではキーワード「オリンピック」に対する結果がグラフ表示されており,灰色の棒グラフは「オリンピック」を含む記事数を,赤線で示されるグラフは burst 度の推移をそれぞれ示している.

このグラフでは,1998 年 2 月(長野),2000 年 9 月(シドニー),2002 年 2 月(ソルトレイク)で blog 記事が有意に密集しており,burst として描かれていることがわかる.

また，2002 年後半部分では「オリンピック」を含む記事数が多くても，全記事数からするとそれほどでもないため，burst となっていないこともわかる．

上で述べたように，burst 度が高くなっているキーワードというのはその期間で注目されているキーワード(ホットキーワード)だと考えることができる．そこで，すべてのキーワードの burst 度を計算し，各月において burst 度の最高値を基準としてキーワードを整列することでホットキーワードリストを作成することができる．本システムでは定期的にこのリスト作成を行っている．

このリストは右図のようにインタフェースから閲覧することが可能である．ここで示しているのは 2002 年 6 月のホットキーワードであるが，ワールドカップサッカーが行われていた期間であることを反映し,「ワールドカップ」,「トルコ」,「イングランド」,「ベッカム」などがホットキーワードであるとされている．

なお，キーワードに関する burst 度の計算の詳細は，[藤木，2004]を参照して欲しい．

3.2.2. 評価表現の検出

blog マイニングの 1 つのアプローチとして,評価(主観)情報の抽出が考えられる．今回はそのための足がかりとして,blog 記事に含まれている評価表現を視覚的にわかりやすく呈示するため，評価表現を自動検出する機能を実装した．

肯定的表現は背景を赤く，否定的表現は背景を青く，肯定か否定か不明だが評価表現になり得る表現は背景を緑にハイライトするようにした．

3.3. お勧め blog の推薦

収集した blog の中から，今注目すべき話題をユーザに提示する機能として，metablog を構築している．metablog とは，面白い blog を発見して紹介するような blog のことを指す． metablog の良い点は，そこにある情報をウォッチしているだけで，今話題の情報を知ることができるという点である．

図 2 は構築した metablog システムのスクリーンショットである．

図 2:metablog

基本的なアイデアは，burst しているキーワードを含む entry を検索し，それらの entry が参照している共通のニュースソースを検索することで，注目するべきニュースを探す．metablog は，movable type を利用して構築している．movable type では，XML-RPC 経由で書き込みが可能であるため，これを使用して blog への書き込みを行っている．

また，一日に一度，お勧めキーワード 5 件が，blog に自動的に書き込まれる．以下に実際のお勧めキーワードの例を示す．

- 2004/1/7
  - 「仏内相」
    →相撲は知的スポーツでないと発言
  - 「震災当日」
  - 「阪神大震災 9」
    →阪神大震災から 9 年
  - 「芥川賞」
  - 「最年少」
    →芥川賞最年少受賞

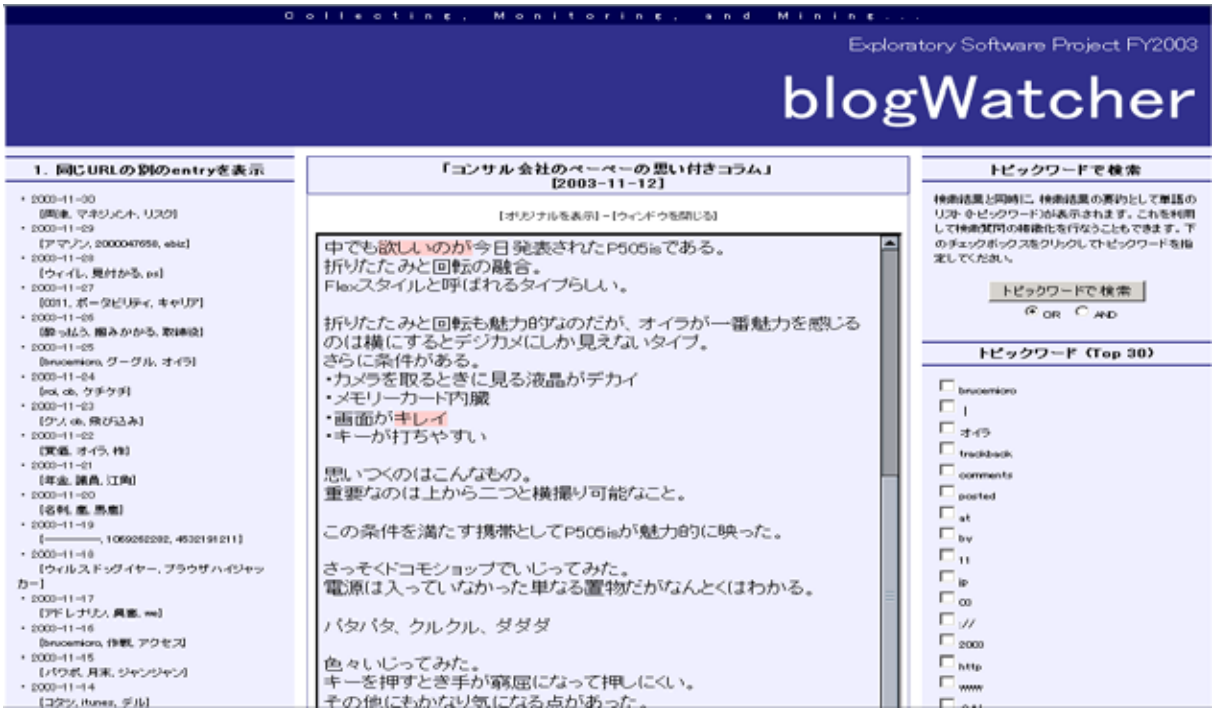
3.4. 評判情報検索

3 節の最初に述べたが，本システムでは，収集した blog ページおよび entry について様々な形での検索が可能である．ここでは，比較的ニーズが高いと思われる，ある対象（キーワード）に関する評判情報（評価表現を含むページあるいは entry）を検索する例を示す．

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なわれる．キーワードとして「505」を入力した結果を図 3 に示す．出力結果は，入力されたキーワードを含む記事 1000 件(検索結果における上位 1000 件)の中から，キーワードが評価対象になっている評価表現を含む文を抽出し表示している．

図 3:評判情報検索

図 3 の 1 つ目の entry を選択すると，下図のように表示される．書き手が対象に対して肯定的な内容を記述していることが読み取れる．



[南野，2004]　南野朋之,鈴木泰裕,藤木稔明,奥村学,blog の自動収集と監視,情報処理学会自然言語処理研究会報告，160-19，pp.129--136，2004.

[Kleinberg,2002]　Kleinberg, J., Bursty and Hierarchical Structure in Streams, Proc. of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1--25, 2002.

[藤木,2004] 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，情報処理学会自然言語処理研究会報告，160-13，pp.85--92，2004.

4. 開発成果の特徴

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．こういう状況を背景に，掲示板を情報源とするテキストマイニングの試みはすでに始まっているが,同様に更新頻度が高く速報性の大きい Web ページである blog を対象としたものはまだ見当たらない．

また，近年 blog は情報源として注目され始めているが，サイトとして運用されており，多数の書き手が書いた記事をまとめて収集できるもののみが対象となっている．このようなサイトは特に日本ではそれほど多くなく，通常の Web ページと変わらず個人が各自書いている blog ページが多数を占めているのが現状であり,サイトとして運用されているもののみを対象としていては，発信されている多くの「生の声」を見過ごしてしまう可能性が大きい．そのため，本プロジェクトでは，WWW 上に散在する blog ページと考えられるページを選択的に収集し,定期的な監視を網羅的に行うことを試みる．

特定の情報源を探索し，選択的に情報を収集するアクティブマイニングの分野でも， 更新頻度が高く速報性の大きい Web ページである blog を対象とし収集するものは見られない．

5. 活用の見通し

今後 DB，プロセスを分散化した上で，システムの運用開始を予定している．

6. 開発者

奥村 学(東京工業大学精密工学研究所　oku@pi.titech.ac.jp)
南野朋之，藤木稔明，鈴木泰裕(東京工業大学総合理工学研究科知能システム科学専攻　{nannno,fujiki,yasu}@lr.pi.titech.ac.jp)
問い合わせ先:blog@lr.pi.titech.ac.jp
URL: http://www.lr.pi.titech.ac.jp/blogwatcher/