3D2-01

Weblogから社会の関心を探る

Finding concerns of people from Weblog articles

福原知宏[*1] Tomohiro FUKUHARA
村山敏泰[*2] Toshihiro MURAYAMA
中川裕志[*3] Hiroshi NAKAGAWA
西田豊明[*4] Toyoaki NISHIDA

[*1] 科学技術振興機構 社会技術研究開発センター
Research Institute of Science and Technology for Society, JST

[*2] エス・ピー・エス・エス株式会社
SPSS Japan Inc.

[*3] 東京大学情報基盤センター 図書館電子化部門
The University of Tokyo

[*4] 京都大学大学院 情報学研究科
Kyoto University

An analysis tool of social concerns using Weblog articles is described. Proposed tool called KANSHIN collects Japanese, Korean, and Chinese Weblog articles, and analyzes them across language boundaries. Some analysis results obtained from KANSHIN are reported.

1. はじめに

近年，Weblogと呼ばれるWebページ管理システムの普及に伴い，国内外を通じて多くの人々がWeblog上で情報発信をしている．Weblog上で発信される情報には個人の身の回りの出来事や社会の出来事に関する意見が含まれている．こうした個人の発信するWeblog記事を大量に収集し分析することで，社会の関心動向を迅速に把握できるようになる．今日の社会において社会の関心動向を知ることは，多くの企業活動や行政サービス，市民生活において重要である．本論文では筆者らの開発しているWeblog記事を用いた社会の関心を解析するシステムについて述べ，本システムを用いて行った分析結果について述べる．

本論文の構成は次のとおりである．2. では提案システムについて述べる．3. では提案システムを用いて得た結果について述べる．4. では本論文のまとめと今後の課題について述べる．

2. Weblog記事を用いた関心解析システム

本節ではシステムへの要求と実装システムについて述べる．

2.1 システムへの要求

筆者らはWeblog記事を用いた関心解析システム：KANSHINの開発を行っている[福原 2005]．システムの開発と試験運用を通じ，現時点におけるシステムへの要求は次のようになった．

(1) Weblog記事の収集に関する要求

1. 記事の大量収集
   社会の関心を把握するため様々なWeblogサイトから大量に記事を収集する．特に複数のPCを用いて効率的に記事を収集する．
2. Weblogサイトごとの収集
   個々のWeblogサイトの関心動向も把握するため，Weblogサイトごとに記事を収集する．
3. 海外の記事の収集
   国内だけでなく海外の記事も収集する．これにより海外の関心動向を知り，国内と海外での関心の比較を行う．

(2) Weblog記事の解析に関する要求

1. グラフによる関心動向の可視化
   関心動向の定量的な把握のため，グラフによる関心動向の可視化を行う．
2. 頻出語の自動検出
   頻出語を自動的に検出する処理が必要である．これにより1日，1週間，1ヶ月単位での関心を把握できる．
3. 言語横断的な解析
   利用者の指定する検索語に対応する訳語を求め，複数の言語の記事を対象として関心動向の比較を行う．
4. サイトごとの解析
   あるサイトがどのような記事やキーワード，ハイパーリンクを含むかを解析し，サイトの特徴を定量化する．

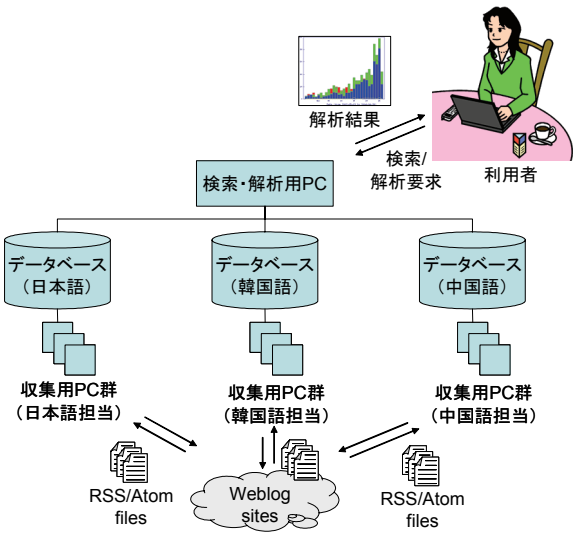

Figure 1. システム構成図

連絡先：福原 知宏，東京大学人工物工学研究センター，
〒277-8568 千葉県柏市柏の葉 5-1-5，
phone: 04-7136-4275，e-mail: fukuhara@race.u-tokyo.ac.jp

5. 他のデータとの比較
   新聞記事や統計データ等，Weblog 以外のデータとの関係を解析する．
6. 共起語を用いた内容の解析
   単語間共起を用い，どのような単語が共起したかを調べる．

2.2 システム構成

Figure 1に実装システムの構成を示す．システムは大量の記事を収集するため，複数の PC から構成される．内訳は記事収集用 PC が 7 台，検索･解析用 PC が 1 台，データベース 3 台の計 11 台である．収集用 PC では日本語サイト担当 PC が 4 台，韓国語と中国語 3 台の計 7 台である．

システムはデータベース中に登録された Weblog サイトを巡回し RSS ファイルもしくは Atom ファイルを収集し，記事と記事に含まれる単語をデータベースに登録する．Weblog サイトの登録は ping サーバや新着 Weblog リストに現れる URL から新規 Weblog サイトを抽出し，1 日平均 17,000 サイト（日中韓合算値）を登録している．

システムは 1 日に 172,800 サイト（日本語サイト），57,600 サイト（韓国語），14,400 サイト（中国語）を巡回し，平均 524,310 件（日本語），272,964 件（韓国語），26,299 件（中国語），計 823,572 件の記事を収集し登録する．

現在[1]，データベースに登録されている記事数は日本語 57,657,230 件，韓国語 10,829,706 件，中国語 3,657,926 件，延べ 72,144,862 件である．また登録されているサイト数は日本語 1,594,082 サイト，韓国語 282,464 サイト，中国語 314,539 サイト，延べ 2,191,085 サイトである．

2.3 機能

本システムは次の機能を有する．

(1) 記事検索機能

利用者の指定した検索語を含む記事を検索する．検索の結果，記事数の推移をグラフで表示する．またサイトごとに記事を表示し，記事数の推移を表示する．

(2) 共起語検索機能

利用者の指定したキーワードと共起する単語を抽出する．

(3) 関心語自動検出機能

単語出現頻度の時間的推移から利用者の指定した期間における頻出語を自動抽出する．これにより利用者は 1 日の話題，1 週間の話題，1 ヶ月の話題を把握できる．

(4) 言語横断型検索

利用者の指定した日本語キーワードを中国語と韓国語に翻訳し，それぞれの言語の記事に対して検索を行う．キーワードの翻訳に当たっては日本語キーワードをWikipedia[2] で検索し，その検索結果から韓国語と中国語へのリンクを辿ることで対訳表現を得る．

3. 分析結果

本システムを用いて行った分析の結果について述べる．

[1] 2006 年 4 月 14 日 16 時 15 分現在

[2] http://ja.wikipedia.org/

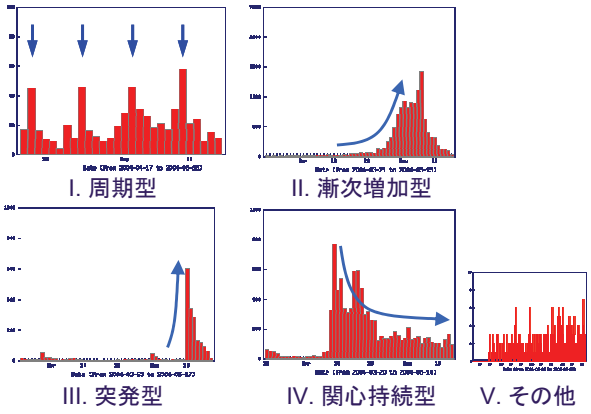

Figure 2. 観測された関心パタン

3.1 関心パタンの分類

我々は日本語 Weblog 記事に含まれる単語の出現頻度の推移から，Weblog に見られる関心パタンを 5 類型に分類した[福原 2005]. Figure 2に関心パタンの一覧を示す．

(1) 周期型

周期性を持つパタンであり，周期的な出来事（テレビ番組や季節の行事）や生活に関連した言葉（“給料日”，“休日”，“家族連れ”など）が該当する．

(2) 漸次増加型

徐々に出現頻度が増加するパタンであり，事前に予期される出来事に関連する言葉（例えば“花粉”，“GW”，“選挙”など）が該当する．

(3) 突発型

急激に出現頻度が増加するパタンであり，重大な事件や事故，災害を示す言葉が該当する．例として“地震”がある．

(4) 関心持続型

人々の関心がある日を境に増加して，その後，持続するパタンである．“ライブドア”や “耐震強度偽装”などが該当する．

(5) その他

(1)から(4)に該当しない言葉であり，“昨日”，“日記”，“事故”など一般的な単語や，調査時点で注目されていない単語（例えば 2006 年 4 月 14 日時点における“酸性雨”など）が該当する．

3.2 実世界データとの関係

Weblog に出現する単語はその時々のニュースや新聞，雑誌，書籍，映画，音楽，天候，行事など実世界の様々な活動の影響を受けている．そこで我々は Weblog 記事中に出現する単語と実世界の現象との関係を調べた．ここでは 2004 年の東京（大手町）の平均気温と相関する単語を日本語の記事から抽出した[Fukuhara2005].

Figure 3に平均気温と相関する単語の時間的変化のグラフを示す．平均気温と正の相関を持つ語の例として，“汗”(0.73)，“虫”(0.82)がある(カッコ内は Pearson の相関係数)．また負の相関を持つ語には，“暖かい”(-0.77)，“寒い”(-0.60)といった単語が見られた．この他，湿度，日照時間，最大瞬間風速，最大瞬間雨量等の気象データとの相関について調べ，気象データと相関性を持つ単語を見出した[Murayama2005].

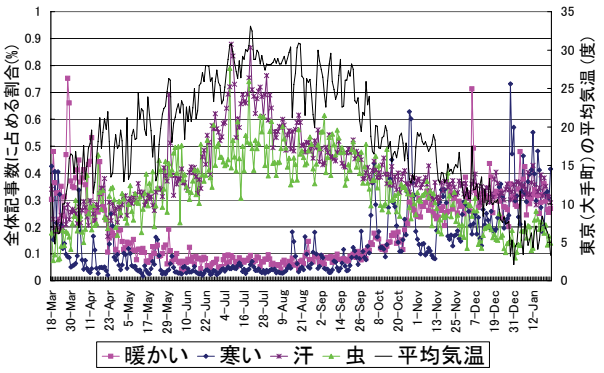

Figure 4. 平均気温と相関する単語の例

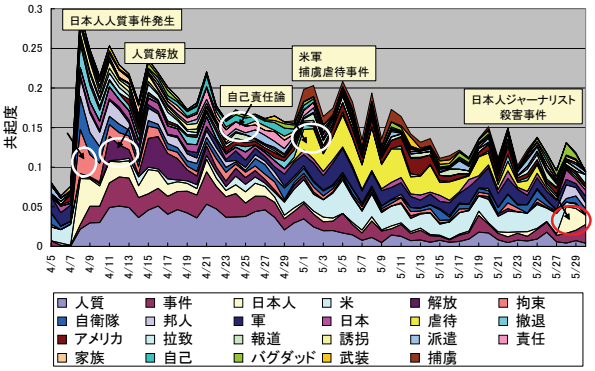

Figure 3. “イラク”と共起した単語の推移 (2004/4/5 から 2004/5/29 まで)

3.3 共起語の時間的変化

ある出来事を示す単語の共起語を時系列で見ることによって，その出来事が人々にどのように捉えられていったかを把握できる．Figure 4は 2004 年 4 月 5 日から 5 月 29 日までの間で，“イラク”と共起した単語の時間的推移を示したものである．Y 軸はそれぞれの日付における各単語の共起の割合である．

Figure 4において 2004 年 4 月は日本人人質事件に関する単語(“日本人”，“拘束”，“解放”，“責任”など)が“イラク”と共起していたが，5 月は米軍による捕虜虐待事件に関する単語(“捕虜”，“虐待”，“軍”など)が共起していたことが分かる．このように共起語の時間的変化を見ることで，ある出来事がどのようにクローズアップされたかを把握できる．

3.4 言語横断型関心解析

今日の社会問題は 1 国だけの問題に留まらず複数の国にまたがって存在する場合が見られる．問題を多面的に捉えるには，複数の国や地域における関心を相互に比較する必要がある．

我々は試験的に日中韓の記事を対象とした言語横断型検索機能を用意した．Figure 5は“鳥インフルエンザ”を含む日中韓の記事数の推移を比較したグラフである．Y 軸はそれぞれの日付において収集した記事数に占めるキーワードを含む記事数の割合である．この図から，(1)“鳥インフルエンザ”に対して関心の集まった時期はそれぞれ異なること，(2)中国語記事における関心は他の言語より高かったことが分かる．

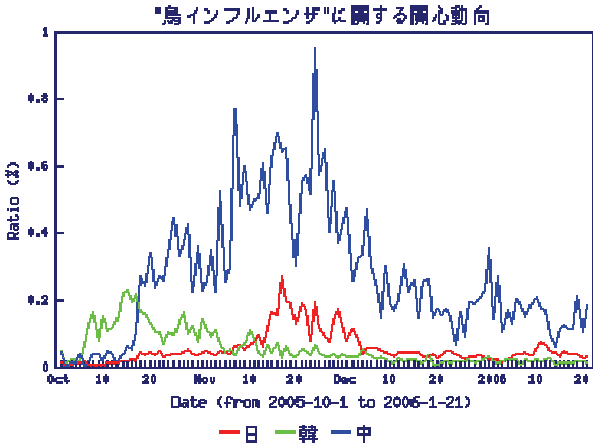

Figure 5. “鳥インフルエンザ”に関する日中韓の関心比較 (2005/10/1 から 2006/1/21 まで)

4. まとめ

本論文では Weblog を用いて社会の関心を解析するシステムについて述べ，本システムを用いて行った分析の結果について述べた．今後，本システムによって得られた関心の妥当性を検証するとともに，(1)収集対象言語の拡張，(2)ハイパーリンク関係の解析，(3) 他の実世界データとの関係の解析などを行う．

参考文献

[福原 2005] 福原, 村山, 中川, 西田: ウェブログ記事を用いた関心解析システム，人工知能学会全国大会, 2C2-04，2005.

[福原 2006] 福原, 中川, 西田: 感情表現と用語のクラスタリングを用いた時系列テキスト集合からの話題検出，人工知能学会全国大会, 2E1-02，2006.

[Fukuhara2005] Fukuhara,T., Murayama,T., and Nishida,T.: Analyzing Concerns of People using Weblog Articles and Real World Temporal Data, WWW2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, Chiba, Japan, May 10th (2005).

[Murayama2005] Murayama,T., Fukuhara,T., and Nishida,T.: Analyzing concerns of people using weblog articles and natural phenomena, in Rajiv Khosla, Robert J. Howlett, Lakhmi C. Jain (Eds), Knowledge-Based Intelligent Information and Engineering Systems: 9th International Conference, Proceedings of KES 2005, Springer LNCS 3683, Part III, pp.855-860 (2005).