2C2-04

ウェブログ記事を用いた関心解析システム

A system for analyzing social and personal concerns from Weblog articles

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

[*1] 科学技術振興機構 社会技術研究システム RISTEX, JST
[*2] 東京大学情報基盤センター 図書館電子化部門 The University of Tokyo
[*3] 京都大学大学院 情報学研究科 Kyoto University

We describe a system for analyzing social and personal concerns of people from Weblog (blog) articles called KANSHIN. The system collects Japanese and Chinese blog articles, and analyzes them. The system finds daily/monthly topics, co-occurrence words, and documents. We describe (1) patterns of social concerns, and (2) change of personal concerns.

1. はじめに

本論文ではウェブログ記事を用いた関心解析システム: KANSHIN と本システムを用いて得られた結果について述べる. 地震, 台風, 津波, SARS, BSE, 鳥インフルエンザなど, 今日, 我々の身の回りには様々な社会問題が存在する. これらの問題は複数の国や地域にまたがる問題でもある. このことから社会問題の理解と解決には, (1)問題の原因に対する自然科学的調査と, (2)問題に対する社会反応や関心の把握といった社会科学的調査の双方を, 国や地域を越えて進める必要がある.

本研究では(2)の観点から, 現在, WWW (World Wide Web) 上で国際的に普及しつつあるウェブログ (ブログ) の記事を大量に収集し解析することで, 社会問題に関する人々の関心を言語横断的に把握することを目指している.

本論文の構成は次の通り. 2 章では提案システムについて述べる. 3 章では社会的関心のパタンについて述べる. 4 章では社会の出来事と個人の関心の関係について述べる. 5 章では関連研究との比較を行う. 6 章では本論文の議論をまとめ, 今後の課題について述べる.

2. ウェブログ記事を用いた関心解析システム: KANSHIN

本システムはブログサイトから RSS 及び Atom ファイルを定期的に収集し, 利用者の求めに応じて記事を解析する. Figure 1 にシステムの概要を示す. システムは日本と中国のブログサイトから収集した記事を記事テーブルに格納すると共に, 記事に対して形態素解析を行い, 名詞と形容詞を抽出して索引テーブルに格納する. 形態素解析システムには茶筌[1] (日本語の場合) と ICTCLAS [Zhang2003][2] (中国語の場合) を用いた. システムは現在 1 日に 2 万件 (日本語), 2 千件 (中国語) の記事を収集しており, これまでに 1 千万件 (日本語), 20 万件 (中国語) の記事を収集している[3].

システムの機能は次の通りである. (1) 記事検索機能: システムは利用者の指定するキーワードを含む記事を検索し, 日ごとの記事数のグラフと記事のリストを提供する. (2) 共起語検索機能: システムは利用者の指定したキーワードと共起する語を検索する. (3)話題検出機能: システムは日ごとの話題や月ごとの話題を検出し, 利用者にメールで配信する.

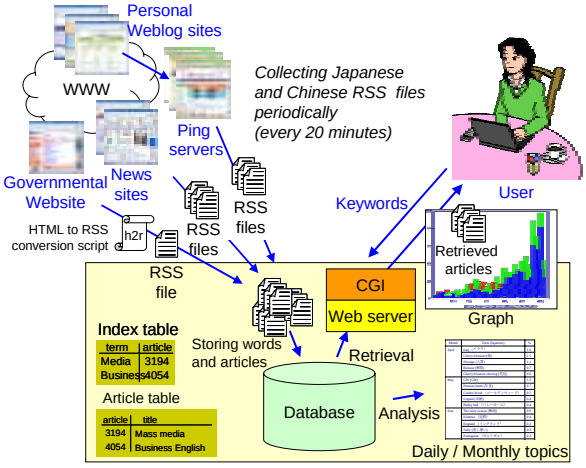

Figure 1. システム概要

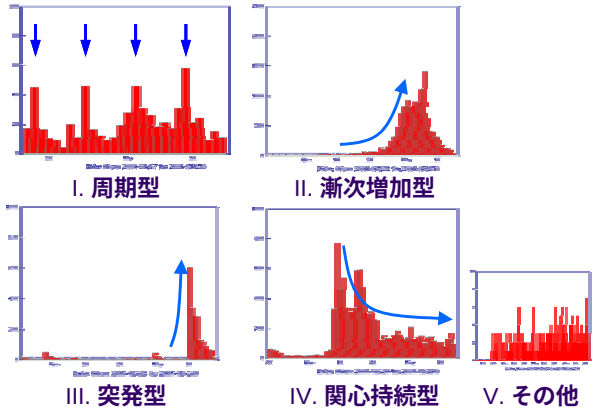

Figure 2. 社会的関心パタンの一覧

3. 社会的関心パタン

本システムを用いて社会的関心を 5 つのパタンに類型化した. Figure 2にパタンの一覧を示す. ここではパタンを (1)周期型, (2)漸次増加型, (3)突発型, (4)関心持続型, (5)その他の 5 パタンに分類した. 以下, 各パタンについて述べる.

[1] http://chsen.naist.jp/ (accessed 2005-04-24)

[2] http://mtgroup.ict.ac.cn/~zhp/ICTCLAS.htm (in Chinese; accessed 2005-03-04)

[3] 2005 年 4 月 24 日正午時点

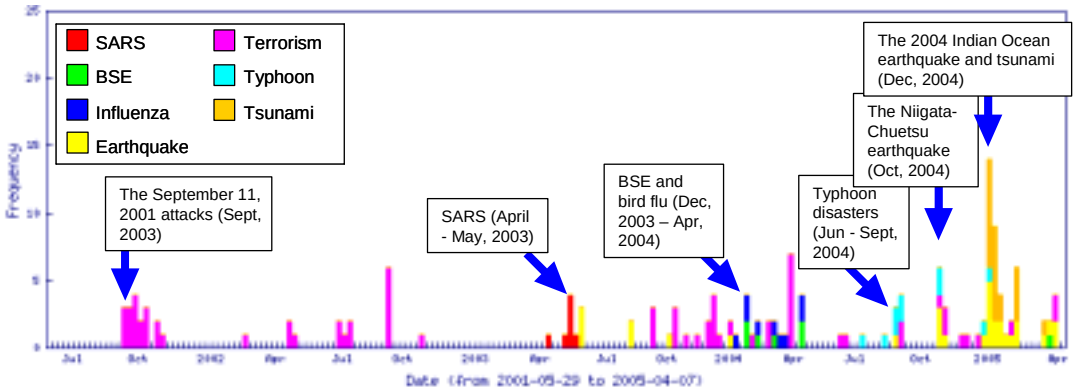

Figure 3. 社会問題に関する小泉首相の関心

3.1 周期型

周期的に記事数が変化する出現するパタンである. このパタンには, (1)テレビの番組名 (“冬のソナタ”, “新撰組”, “エンタの神様”など) や(2)生活に関連する周期的な出来事 (“週末”, “BBQ”, “家族連れ”, “給料日”, “出社”, “授業”など) が該当する. 長期的には“夏休み”, “オリンピック”なども該当する.

3.2 漸次増加型

徐々に記事数が増えるパタンである. Figure 2のパタン II のグラフは 2004 年 4 月から 5 月前半にかけての“GW”を含む記事数の推移である. この図に見られるように, 記事数は連休に向かって徐々に増加し, 連休が過ぎると共に徐々に減少している. このことから人々が事前にある出来事の存在を把握しており, その出来事について強い関心を抱いている場合に漸次増加型パタンが生じると考えられる. “台風”, “選挙”, “夏”, “クリスマス”などはこのパタンに該当する.

3.3 突発型

急激に記事数が増加するパタンである. このパタンは, ある出来事に対して人々がその存在を予期しておらず, かつその出来事に強い関心を抱いている場合に生じる. Figure 2のパタン III のグラフは 2004 年 4 月から 5 月前半における“Winny”を含む記事数の推移である. グラフ中のピークは 2004 年 5 月 10 日の Winny 開発者逮捕を受けた人々の反応である. このグラフから, 人々が開発者逮捕について強い関心を抱いていたことが分かる. このパタンを代表する他の語には“地震”がある.

3.4 関心持続型

ある出来事が発生した後に, 人々の関心が持続しているパタンである. Figure 2のパタン IV のグラフは“イラク”を含む記事数の推移である. 2004 年 4 月に発生したイラク人質事件により, 人々の関心は急激に高まり, 持続した関心となっていることが分かる. このパタンに該当する他の語には“ニッポン放送”がある.

3.5 その他

上に挙げたパタンのいずれにも該当しないパタンである. これに該当する語は, (1)日常的に使われている語 (“日記”, “今日”, “ブログ”など) や, (2)人々が現時点において関心を示していない語 (“食糧問題”, “水不足”, “海洋汚染”) などである. ただし後者に属する語でも一旦問題が顕在化すると突発型や関心持続型に移行する例も見られる[1].

[1] 例えば“原発”は 2004 年 8 月に発生した関西電力美浜原子力発電所事故を受けて突発型を示した.

4. 社会問題に関する個人的関心の分析

社会問題に対する個人の関心に関する予備実験として, 小泉首相のメールマガジン[2]を用いた分析を行った. Figure 3に分析結果を示す. x 軸は日付を, y 軸は単語の出現頻度である. Figure 3に見られるように小泉首相の関心はその時々の社会の出来事に応じて高まる半面, “テロ”を除いてそれぞれの問題に対する関心が継続していない事が分かる.

5. 関連研究

南野らはウェブ日記とウェブログを対象とした関心解析システム blogWatcher を提案している[Nanno2004]. また Glance らは英語のブログを対象とした関心解析システム BlogPulse を提案している[Glance2004]. 本研究とこれらのシステムの違いは, (1)個人単位での関心分析が可能であること, (2)複数の言語における関心解析にある. (2)に関しては現在, 手動で日本と中国の関心比較を行っており, 今後, 異なる言語間で同じ概念を表すキーワードを自動検出し, そのキーワードを用いて関心の言語間比較を行う機能の実装を検討している.

6. まとめと今後の課題

本論文ではブログ記事を用いた関心解析システム: KANSHIN を提案し, 本システムを用いた分析結果について述べた. 分析の結果, 5 つの社会的関心のパタンを見出した. また, 小泉首相の関心がその時々の社会の出来事に応じて高まることが判明した. 本研究の今後の課題は, (1)システムの扱う言語の拡張と, (2)関心の言語間比較である.

参考文献

[Zhang2003]Zhang, H.P., Yu, H.K., Xiong, D.Y., and Liu, Q. HHMM-based Chinese lexical analyzer ICTCLAS, In Proceedings of Second SIGHAN Workshop on Chinese Language Processing, pages 184–187, 2003.

[Glance2004]Glance, N., Hurst, M., and Tomokiyo, T., BlogPulse: Automated trend discovery for Weblogs. In WWW 2004 Workshop on the Weblogging Ecosystem, 2004.

[Nanno2004] Nanno, T., Suzuki, Y., Fujiki, T., and Okumura, M. Automatic collection and monitoring of Japanese Weblogs. In WWW 2004 Workshop on the Weblogging Ecosystem, 2004.

[2] http://www.kantei.go.jp/jp/m-magazine/ (accessed 2005-04-24)