様式 C-19

科学研究費補助金研究成果報告書

平成２２年　５月３１日現在

研究種目：若手研究(B)
研究期間：2008 ～ 2009
課題番号：20700222
研究課題名（和文）　ソーシャルメディアとマスメディアの相互影響分析システムの構築
研究課題名（英文）　Development of Analysis System for Mutual Influence on Social Media and Mass Media

研究代表者
石田　和成（KAZUNARI ISHIDA）
広島工業大学　情報学部　准教授
研究者番号：20303026

研究成果の概要（和文）：ブログ，ニュース，スパムのキーワード時系列データにおいて，自己相関にもとづき，各情報源の周期的話題の分析を行った．そのため，独自で継続的に収集しているブログとニュースのデータを用いた．また，独自に開発したスパム分離手法を開発し，収集したブログからスパムの分離を行った．これら時系列データの違いを調べるために，自己相関にもとづくキーワードの文書頻度の基本周期系列抽出アルゴリズムを開発し，システム構築を行った．このシステムを用いて周期の分布や，7 日周期および 365 日周期のキーワードの抽出を行った．その結果，ブログは毎週のテレビ番組や週末の趣味や年中行事，ニュースは政府や経済，スパムはメルマガやアフィリエイトの話題が多いことが分かった．

研究成果の概要（英文）：Time-series data of keywords within blogs, news, and spam is analyzed in terms of auto-correlation to find periodic topics in these information sources. The information is collected from Japanese blog sites and news sites. Spam blogs are then separated from legitimate blogs using a spam filtering system. To find differences among the three sources, an analysis system is developed to find periodic topics based on auto-correlation. Employing this system, distribution periods of keywords within each information source, weekly keywords, and yearly keywords are extracted. In terms of distribution and keywords, characteristics of information sources are illustrated. According to the results, periodic blog topics are TV programs, hobbies, and social events. Periodic news topics are political and economical events. Periodic topics in spam are automatically copied-and-pasted email newsletters and affiliates.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
平成２０年度	700,000	210,000	910,000
平成２１年度	500,000	150,000	650,000
年度			
年度			
年度			
総　計	1,200,000	360,000	1,560,000

研究分野：総合領域，情報学
科研費の分科・細目：図書館情報学・人文社会情報学
キーワード：情報組織化，テキストマイニング，グラフマイニング

１．研究開始当初の背景

(1) ソーシャルメディアにおける話題抽出に関して，内田，柴田(2006)はトラックバックやコメントといったブロガー(ブログの筆者)

の相互関係にもとづくコミュニティ抽出を行い，各コミュニティの話題抽出を試みた．しかし，コメントやトラックバックは，ブロガーの人的つながりを示すもので，特定の話題の共有を必ずしも保障しない．そのため，ソーシャルメディアの話題抽出には，特徴的な話題を抽出する仕組みを考案する必要がある．

(2)マスメディアのニュース記事における話題抽出に関して，平田，大囿，新谷（2007）は，記事が掲載される頻度に応じて，記事の集まりを分割する時間幅を変更し，分割された記事をクラスタリングすることにより，話題抽出を行うシステムを開発している．しかし，この研究ではマスメディアと比べ記事の内容にばらつきが大きいソーシャルメディアにおける話題抽出は対象としていない．

(3)ソーシャルメディアとマスメディアに関する研究として，Adamic and Glance（2005）は，支持政党の違いによるブロガーの特徴の違いを分析し，アメリカにおける２つの政党支持グループにブロガーを分け，ニュースサイトへのリンクや，グループ内，グループ間でのリンクの張り方の傾向を分析した．しかしこの研究は２つのメディアの相違や相互影響の分析は目的としていない．

これらの国内，国外の研究動向を踏まえ，本研究では，ソーシャルメディアとマスメディアの相互影響分析システムを構築する．このシステムの構成要素として，これまでに開発したブログやオンラインニュース記事の収集システムの拡張を行うとともに，特定の話題を共有したブログコミュニティの抽出手法やスパムブログの分離手法の開発を行った．

２．研究の目的

本研究の目的は，市民のブログ記事で構成されるソーシャルメディアと，専門家のニュース記事で構成されるマスメディアとの間の相違や，相互に与える影響について定量的に分析するためのシステムを構築することである．そのため，システムを構成する要素技術として,ブログとオンラインニュースの収集,蓄積，スパムブログ分離の仕組み，特徴的なトピックの抽出手法を開発する．

３．研究の方法

(1)ブログ,オンラインニュースの収集蓄積システムの開発を行う．ソーシャルメディアとマスメディアの相違や相互影響を網羅的に分析するために，大規模なデータを効率的に扱うことができるシステムの設計，構築，運用を行う．

(2)スパムブログ分離システムの開発を行う．これは，スパムブログとスパムキーワードのの共起クラスターと，連鎖的なスパム抽出手法にもとづき開発する．このシステムを用いて，ソーシャルメディアにおけるノイズであるスパムを分離することにより，精度の高い分析を行うことができる．

(3)ソーシャルメディアとマスメディアの相互影響分析システムの開発する．このシステムにもとづき，ブログ，ニュース，スパムの特徴を調査する．このシステムでは，それぞれの情報源における特徴的なトピックを抽出するために，自己相関にもとづきキーワード出現頻度の基本周期系列を抽出するアルゴリズムを用いる．そして，この周期検出アルゴリズムを用いて，各情報源の周期分布や，7日周期および 365 日周期のキーワードを抽出し，ブログ，ニュース，スパムの特徴を分析する．

４．研究成果

(1)話題共有コミュニティ抽出システムの開発のために，Wikipedia にもとづくキーワードの表記ゆれおよび同義語抽出手法の開発を行った．ソーシャルメディアにおける話題を表すキーワードには多様な表記ゆれ，同義語が存在するため，話題の特定が難しい．そのため，Wikipedia におけるエイリアス（図１）もとづくキーワードの表記ゆれ情報の抽出手法を開発し,抽出条件の調査を行った(図２)．人手で編集された表記統合辞書との比較を行った．その結果，Wikipedia と表記統合辞書は補完的であり，Wikipedia から抽出されたこの用語間関係は，キーワードの多様性の高いブログのデータにおいて，流行や意見を抽出するための，キーワードデータベースとして利用できる可能性があることが分かった．

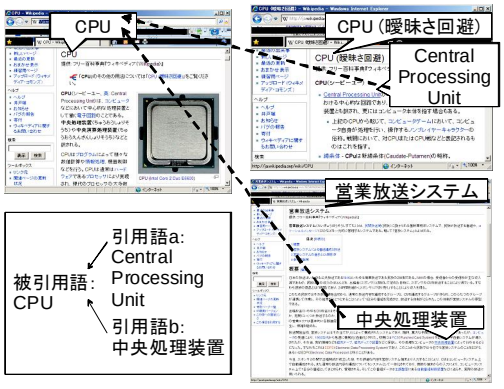

図１：エイリアスにもとづく用語間関係

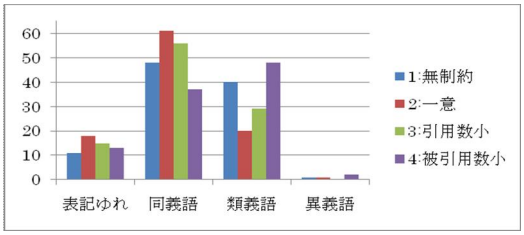

図２：抽出された用語集合の分布

(2) 更新ブログのデータからスパムブログを分離する手法を開発した．ブログは意見や評判などの情報源として注目されているが，多くのスパムブログの存在が問題となっている．スパムブログは，他のブログやニュースなどを断片的にコピーした記事や，露出機会増大のためにマルチポストされる記事が多く存在する．そのため，ブログとキーワードの二部グラフにおいて，スパムブログとスパムワードは，大規模なスパムクラスターを形成する不偏的特徴がある．本研究では，このクラスターをスパムシード（図３）として利用し，連鎖的にスパムブログとスパムワードを相互抽出する（図４）．約６か月間のデータにもとづく予備実験と，１ 日のデータを用いた詳細な調査により，最高 95%の精度でスパム分離ができる標準的なパラメータセット（表１）の検討を行った．

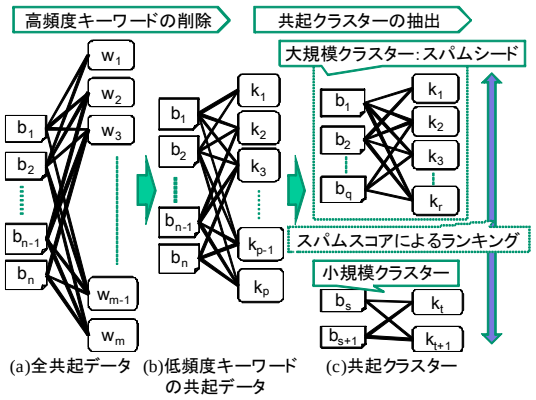

図３：スパムシードの抽出

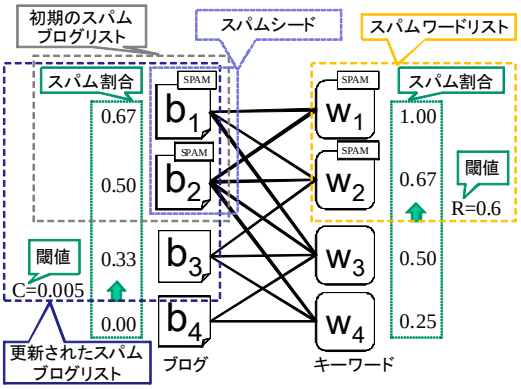

図４：連鎖的スパムブログ/ワードの抽出

表１：標準パラメータ

S	W	R	C	F
0.20	100	0.60	0.005	0.50

(3) ソーシャルメディアとマスメディアの相互影響分析システムの開発のため，更新データ収集システムとスパム分離システムにもとづき，ブログ，ニュース，スパムのキーワード時系列データにおける周期的話題の分析を行った．これら時系列データの違いを調べるために，自己相関にもとづくキーワードの文書頻度の基本周期系列抽出アルゴリズムを開発した．具体的には，個々のキーワード時系列における自己相関のピークを検出し，基本周期の系列を抽出するアルゴリズムを開発した．図５はキーワードの時系列データ，図６はその自己相関の例である．

キーワード数の時系列： $\{y_1, y_2, \cdots, y_n\}$

標本平均： $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y_t$

標本自己共分散関数：

$$\hat{C}_k = \frac{1}{n}\sum_{t=k+1}^{n}(y_t - \bar{y})(y_{t-k} - \bar{y})$$

標本自己相関関数： $\hat{R}_k = \frac{\hat{C}_k}{\hat{C}_0}$

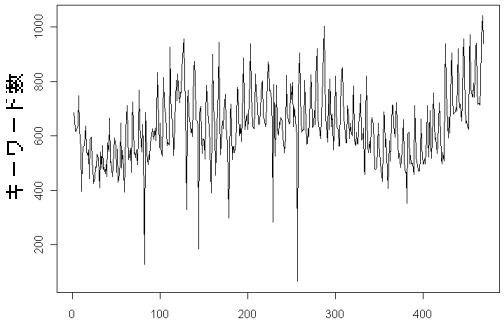

図５：時系列データ

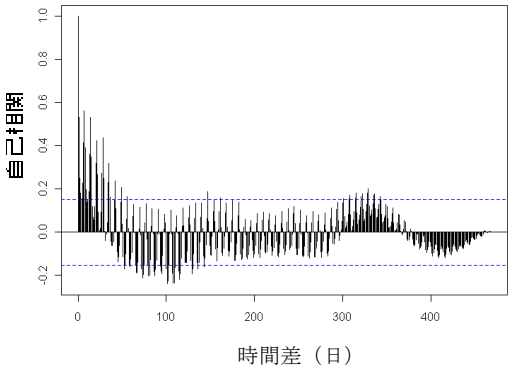

図６：自己相関

また，各キーワードについて，自己相関ピークの特徴量として，標本自己相関係数の隣接差分積 z_k を定義した．

$$z_k = (R_{k+1} - R_k)(R_k - R_{k-1})$$

自己相関における周期のピークでは，隣接する差分の符号が，正から負へ，あるいは，負から正へ，反転するため，隣接差分積 z_k が負となる周期 x をピークと考えることができる．

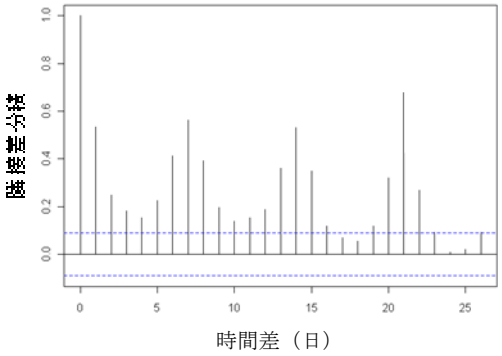

図 7 ：隣接差分積

このアルゴリズムを用いて周期の分布や,7 日周期および 365 日周期のキーワードの抽出を行った．その結果，情報源におけるトピックの共通点や相違点の検出に役立つことが分かった．具体的な自動検出された相違点として，ブログは毎週のテレビ番組や週末の趣味や年中行事，ニュースは政府や経済，スパムはメルマガやアフィリエイトの話題が多いことが分かった．これらの手法にもとづく分析システムは，周期的な話題の自動検出を行うため，定期的に行われる趣味や行事に関連した商品，サービスの販売促進に活用できる．

５．主な発表論文等
（研究代表者，研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 1 件）

1. Kazunari Ishida, ″Spam Blog Filtering with Bipartite Graph Clustering and Mutual Detection between Spam Blogs and Words,″ Journal of Digital Information Management, Refereed, Vol.8, Issue2, 2010, pp.108-116.

〔学会発表〕（計 5 件）

1. 石田　和成, ″ソーシャルメディア分析のための集合知にもとづくキーワードデータベースの自動構築″, 情報システム学会 第5回 全国大会･研究発表大会, 2010 年 1 月 22 日, 電気通信大学.

2. 石田　和成, ″Wikipediaにもとづくキーワード表記ゆれおよび同義語の抽出″, 情報システム学会 第 5 回 全国大会・研究発表大会, 2009 年 12 月 6 日, 青山学院大学青山キャンパス.

3. 石田　和成, ″オンラインメディアにおける周期的トピックの抽出″, 経営情報学会 2009 年秋季全国研究発表大会, 2009 年 11 月 14～15 日, 県立広島大学広島キャンパス.

4. Kazunari Ishida, ″Mutual Detection between Spam Blogs and Keywords Based on Co-occurrence Cluster Seed,″ First International Conference on 'Networked Digital Technologies (NDT 2009), July 28 - 31, 2009, SB Technical University, Ostrava, Czech Republic.

5. 石田　和成, ″キーワードの時系列データにもとづく，ブログ，ニュース，スパムの分析″,第 15 回 Web インテリジェンスとインタラクション研究会, 2009 年 7 月 4～5 日, 広島市立大学.

６．研究組織
(1)研究代表者
石田　和成（KAZUNARI ISHIDA）
広島工業大学　情報学部　准教授
研究者番号：20700222