Webマイニングによるホットトピックモニタリング

西野 文人†, 津田 宏†, 牛 軍佼‡, 黄 萱菁‡, 呉 立徳‡
†富士通研究所, ‡復旦大学
{Nishino,HTsuda}@jp.fujitsu.com, niu_jy@sina.com, {xjhuang,ldwu}@fudan.edu.cn

1 はじめに

インターネットを使っての情報発信が定着し，固定的な情報だけでなく，新聞社等のニュース記事や企業からの技術や製品などの発表記事などがすばやくWeb上で公開され，またそれらに関する批評なども掲示板や評論・日記サイトなどに掲載されるようになっている．このようにインターネットの影響がますます大きくなっている今日，ある分野における世の中の動向を知りたいとか，大量に発信される情報の洪水の中でおさえておくべき情報を知りたいとか，消費者クレームや噂からの信用失墜や株価急落を回避・予知したい[1]というようなインターネット上のホットトピックスのモニタリングに対する要望が高まっている．そこで我々はWeb情報をマイニングすることによりホットトピックスをモニタリングするシステムについて検討している．本稿では，Webページ間のリンク解析（Webマイニング）結果に基づいて，被リンク数やリンク元のページの権威度を利用したWebページの注目度を求める手法を紹介するとともに，この手法を拡張したニュース記事の注目度を求める手法についても述べる．

2 Webマイニングとホットトピックス

自然言語処理の分野ではこれまで新聞などのきっちりしたテキストを対象にして様々な研究開発が進められてきたが，近年のインターネットの普及により，Webを対象としたマイニングにも関心が高まってきている．Webからの情報マイニングとしては，(1) 掲示板の発言に対するアクセスカウンタや検索ログなどの利用状況を使用して分析するもの，(2) 構造（ハイパーリンク）を分析するもの，(3) コンテンツ自身を分析するもの，というように3つに分類することができる[1].

我々の目指すものは，ユーザから指定された分野におけるホットトピックスを見つけることであるが，実際には一口でホットトピックスのモニタリングと言っても「現在のホットな話題をみつける」，「ホットな話題になりつつあるものをいち早くみつける」，「日々発生しているイベントに対して議論沸騰しそうなネタをみつける」というように分類することができる．

このような中，ホットトピックスを見つける試みとしては，WWW検索ログの検索キーワードをグルーピングすることでホットな話題を見つけるもの[2]，時間情報を持つテキストの分類に基づいてホットな話題を見つけるもの[3]，テキスト中の「有名な」とか「注目されている」などの評価表現を手がかりにしてホットになりつつある話題を見つけようとするもの[4]などがある．今回の我々の試みは，構造およびコンテンツ自身を利用することで，注目されているWebページや記事を見つけようというものである．

3 ホットなWebページの発見

それぞれのWebページがどれだけ注目されているかは，どれだけの人がそのWebページを閲覧したかを調べればよい．しかし残念ながら，このアクセス数は公開されているとは限らない．そこで，ハイパーリンクを手がかりとして人気度の高いページ（多くの人が注目しているページ）をランキングする手法がGoogleのPageRank[5]をはじめとしていろいろ提案されている．PageRankの考え

[1] 東芝クレーマー事件や浜崎あゆみの年末コンサート差別発言問題などがある

方は，「多くのページからリンクされているページは人気が高く，人気があるページからリンクされているページも人気が高い，という仮定により再帰的計算による不動点として人気度をとらえている」というものであるが，我々はそこにバラエティ数と新鮮度とを持ち込んで以下の観点で注目度を求めている．

1) 被リンク数
   多くのページからリンクされているページは，注目度の高いページであり，リンクをたどってアクセスされる機会も多くなると考えられる．
2) バラエティ数
   観点の違う様々なサイトからリンクされているページは注目度の高いページである．
3) 権威度
   権威の高いページ（多くの人からアクセスされるページ）からリンクされているページは注目度の高いページである．
4) 新鮮度
   新しい情報を記載しているページは注目度の高いページである．

我々はクローラで収集起点を変えながら毎日収集した約 1.3 億の URL に対して，上記の考え方に基づいてページの注目度（人気度）を求めている [6]．図 1は時間を追っての Web ページの注目度ランキング値の推移をプロットした例である．ホットな Web ページの発見は注目度ランキング推移を線形回帰して，立ち上がりつつある（注目度急上昇）サイトを調べている．例えば，go.jp サイト関連では，2001.7 の小泉内閣メールマガジン，2001.10 の日本科学未来館，2001.11 の厚生労働省「牛海綿状脳症 (BSE) 関係」が注目度急上昇したサイトである．

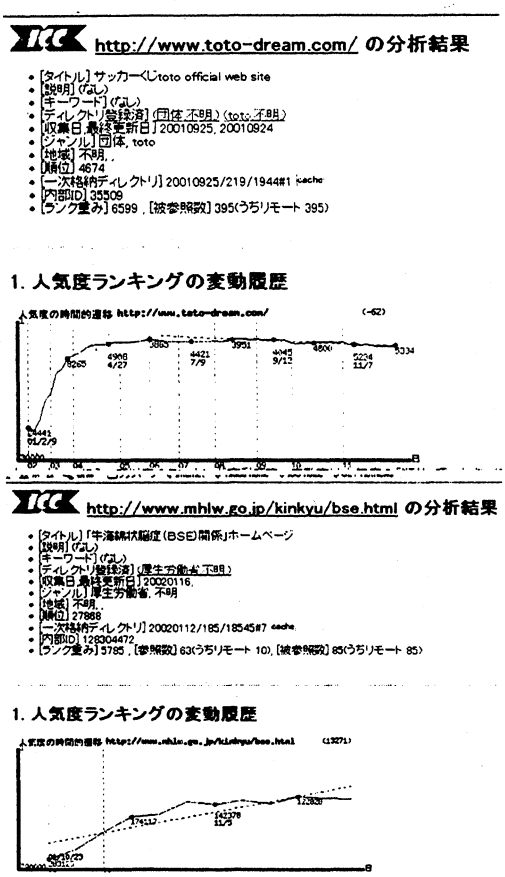

図 1: Web ページ注目度ランキング（上はサッカーくじ (toto, スポーツ振興くじ)HP，下は厚生労働省狂牛病 HP）

4 ホットなニュースの発見

ホットトピックスのモニタリングとして Web ページを単位としていたのでは粒度が粗い．ここではホット Web ページのランキングの延長としてニュース単位でのホットトピックスモニタリングを検討する．まず Web ページを単位としたときと比べての特徴としては以下のものがあげられる．

1) 1 ページに一つのニュース（出来事）だけが記述されているわけではない．
   一つのページに多くのニュースが記述されているページは多い．Web ページの注目度を求める場合には，個々のニュースの内容ではなく，そのページ（サイト）がどのような情報を扱っているかに興味があったので，リンク解析の際にこの点を特に気にする必要はないが，ニュースの注目度を求める場合にはどのニュースが注目されているかを気にしなくてはならない．
2) 本来の情報源は別にあることも多い．
   Web ページの注目度とは Web ページ自身が情報源そのものであったわけだが，一般のニュースは世の中の出来事（事件や発表，

発言，行動，変化など）が情報源である．この情報源（1次情報源）に対して，新聞記事などが，文章で出来事を説明することで参照し，これが2次情報源となり，以降はこの2次情報源が参照されることになる．

3) 関連ページ間でリンクがないことも多い．
注目の対象がページではなくニュース自身であるので，そのニュースの参照はニュースが記述されているページへのリンクではなく，言語表現で参照されることも多い．また2次情報源（各社の新聞記事など）の間でのリンクもないことが多い．

このような状況から，我々はホットなWebページの発見手法に，内容の関連性情報を加味した手法を提案する．まず1次情報源は陽にWebページ化されていなくても仮想のページが存在すると考え，2次情報源は1次情報源の仮想ページをリンクしていると考える．ニュース記事間の関連性を見つけて有向グラフで表現する手法が提案されている[7]が，我々もまた，Webページの内容の類似度によってページ間に仮想リンクが存在すると考える．このような仮想ページと仮想リンクを想定することで，実際のリンクと仮想リンクとを合わせて，Webページの注目度を求めたリンク解析の手法が活用できることになる．

Semantic Web[8]の世界では，これらはメタ情報として記述されるものであるが，現状では，ある記事に対して関連Webページの検索を行い，ニュース記事間の関連性にしたがって仮想リンクをはることになる．ニュース記事間の関連性としては，本来の同一事象を扱った記事だけでなくても，多くの人が関心のある企業の動向であるとか，多くの人が関心のあるテーマであるという関連性も考慮すべきと考え，元の記事から重要キーワード（その記事のテーマを示すキーワード）を抽出してそれをベースに検索システムで検索を行い，tf・idfに基づいて記事間の関連性を判定している．

5 システム評価手法

現在，特定のニュース記事群に対して，その記事の注目度を与えるシステムを構築中であるが，システムをどのように評価すべきかが課題としてある．

5.1 注目度スコアによる評価

我々が現在採用している評価手法は以下のものである．

評価手法

1) 何人かの被験者に特定のニュース源から特定の期間のニュース記事を読んでもらう．
2) 被験者には読んだ記事の中で関心をもった記事に興味度の点数を与えてもらう．
3) システムが出力した注目ニュースと被験者が与えた興味度の点数との相関性で評価する．

この評価手法の課題は，興味度とは何なのかを明確にすることである．一つの方法は，興味度を認知度，すなわちその記事に書いてあることを既に知っていたかどうかでスコアを付与するものである．これは多くの人が知っているニュースは注目度が高いホットなニュースであると考えるものである．もう一つの方法は，記事の内容に対する関心の度合のスコアをつけてもらうものであり，未来のホット話題性の予測につなげるものである．

5.2 評価セットの作成

ある記事群に対する注目度を得るために，人手による注目度付与を行った．この実験ではまず人が注目しているニュースにマーキングをし，それを集計することによりニュースの注目度を求めた．具体的には題材としてACM technewsを使用した．ACM technewsは1回の配信で通常18ないし19の記事があるが，6回の配信分に関して，9人の情報処理技術専門家が数週間前のACM technewsの各記事に対してその注目度に応じて0から3点までのスコアを与えた．ここでは，題材が専門分野であり，それほどニュース性の高い記事でないこともあり，必然的に関心の度合といった意味づけでのスコア付けとなっている．その結果の一部（1回の配信分）を表1に示す．ここで各行は記事，各列は被験者を示し，マイナス記号は被験者がその記事に興味を示さなかった，数字は興味を示した（大きい数字の方が強い興味）ことを示している．

5.3 評価セットの結果について

統計的検定により，この注目度スコアがランダムであることは否決されるが，注目される記事にかな

記事									
1	2	-	3	3	3	3	3	3	-
2	-	-	-	3	2	2	-	2	-
3	2	2	-	-	-	-	-	-	3
4	-	-	-	-	3	-	-	-	-
5	-	3	-	3	-	3	3	-	3
6	-	-	-	-	-	-	-	-	-
7	-	3	3	-	-	-	2	3	-
8	-	-	-	-	-	-	-	-	-
9	-	-	-	-	-	-	-	-	-
1-	-	-	3	-	2	-	-	-	2
11	2	2	-	-	-	-	2	-	-
12	-	-	-	-	-	-	-	-	-
13	2	-	-	-	-	1	-	-	-
14	-	2	3	-	-	1	2	-	-
15	-	-	-	2	3	-	-	3	2
16	-	-	-	-	-	-	-	-	-
17	-	-	-	-	-	-	-	-	-
18	-	-	-	-	-	-	-	-	-
19	2	-	3	2	-	-	-	3	2

表 1: 注目度スコア

りのバラツキがあることがわかる．現在各記事のスコア合計とシステムのスコア値のランクングに基づいて評価を行っているが，被験者数を増やすか別の評価手法の導入を検討するかが必要と考えている．

6 今後の課題

今後様々なニュースサイトのニュース記事や掲示板の発言などの注目度を常にモニタリングするものを作成していく予定であるが，ホットトピックスが本当に抽出できているのかを考えたとき，いくつかの問題点が見えてきている．まずは本当のブームの兆しよりは少し遅れて発見されるという時間的な問題，次に小さなブームが掴みにくいという問題，そしてホットな話題ではなくなったことを認識しにくいという問題である．このような問題が起こる主要な原因は，リンクを張って管理するという作業がコストが高いことにある．このような問題に関しては，リンクを張るコストを下げる，フロー系情報の重要視，意味的な解釈の導入などの対処がある．リンク付与のコストに関しては，我々は閲覧したページが注目に値すると感じた時に簡単な操作で特定のグループ（どのカテゴリの グループかを指定することができる）内の人に知らせる仕組みを用意しており，このようなツールを導入しているグループ内では，グループ全体の興味に焦点をあてたホットトピックスの素早い把握が可能と考えている．また，ニュースや掲示板などの更新度の高いフロー系の情報からのリンクを優先させたり，あるいはリンク関係になった日時の保存やリンクの意図の分類（ゴミリンクの排除を含む）をすることによる効率の良いリンク解析も有効な解決策と考えている．

7 おわりに

我々はリンク解析に基づくホットなWebページを求める（Webページの注目度をランキングする）システムと，検索によって関連記事を探し関連記事間にリンクを張るという考え方でホットなニュースを求めるシステムを作成している．このホットトピックスマイニングは，新聞記事や各企業のプレスリリース，掲示板といったもののモニタリングに役に立つと考えている．また，一言でホットトピックスと言っても，いろいろな観点があるので，それらについても検討していく必要があると考えている．

ホットトピックスがきっちり取り出されたかどうかを評価することは非常に困難である．現在は各記事に対して注目度を人手で付与したものと，システムが与えた注目度のスコアとの比較をすることによって評価しようとしているが，より精密な評価手法を考えていくことも必要と考えている．

謝辞 本研究を進めるにあたり協力をいただいた富士通研究開発中心有限公司の石崎洋之氏，徐国偉氏，復旦大学の呉研究室の学生諸氏に感謝の意を表します．

参考文献

[1] Kosala, R. and Blockeel, H.: Web Mining Research: A Survey, *ACM SIGKDD*, Vol. 2, No. 1, pp. 1–15 (2000).

[2] 大久保雅且, 杉崎正之, 井上孝史, 田中一男： WWW検索ログに基づくトレンド情報の抽出について, 情処研報, DD7-4 (1997).

[3] 杉崎正之, 井上孝史, 大久保雅且, 田中一男：情報分類を用いたトレンド・アウェアネスの支援, 情処研報, DD6-2 (1997).

[4] 落谷亮, 西野文人：評価表現を利用した高認知度情報の抽出, 言語処理学会第 6 回年次大会, pp. 308–311 (2000).

[5] Brin, S. and Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine, in *7th World Wide Web Conference(WWW7)* (1998).

[6] 津田宏, 鵜飼孝典, 三末和男： Web ディレクトリのためのページメタデータの自動付与の試み, 情報学シンポジウム 2002 (2002).

[7] Uramoto, N. and Takeda, K.: A Method for Relating Multiple Newspaper Articles by Using Graphs, and Its Application to Webcasting, in *COLING-ACL98, Vol.2*, pp. 1307–1313 (1998).

[8] W3C, : *Semantic Web Activity*, http://www.w3.org/2001/sw.