情 報 学 基 礎 69－4
（２００２．１１．６）

検索ログからの話題抽出に向けて
－サイト種別の自動判定の試み－

内野寛治　　西野文人
富士通研究所　富士通研究開発中心

インターネット上の雑多な情報の中から注目されつつ話題を抽出する技術は、トレンド把握や風説流布の監視などのためにも重要な技術である。我々はリンク構造に着目してホットトピックの抽出を試みたが、リンクが張られる時間と実際に話題が立ち上がる時間のずれが課題であった。そこで我々は時間のずれが少ないと思われる検索ログを使ったアプローチを現在進めているが、その前処理として大量な検索ログの中から対象ログを絞り込む技術は解析の精度を上げるためにも重要である。本稿では、(時間、検索キーワード、飛び先URL)という３つ組のデータを解析することによって、サイト種別を自動的に判定し、話題の抽出対象となるログを絞り込む手法について提案し、実際の検索ログを使ってその有効性を確認する。

Towards the Extraction of Topic by Analyzing Search Log
－Automatic detection of WWW-Site type－

Kanji UCHINO　Fumihito NISHINO
Fujitsu Laboratories Ltd.　Fujitsu R&D CENTER CO.,Ltd.

The extracting topic from the miscellaneous information on the Internet is so important for the trend grasping, the surveillance of rumor spread, etc. Although we tried to extract the hot topic by analyzing link structure, we found the time gap between the num of links and rising of real topic. Then, we have been trying another approach, which is using the search log data, but it is very difficult to filter the noise from the large log data. In this paper, in order to filter the noise, we propose the URL-site classification method by analyzing the log data, which is the set of (access time, search query, referenced URL).

1. はじめに

インターネット上にある情報の洪水の中から注目されつつある情報を浮かび上がらせる技術、すなわち Web 上の雑多な情報を対象としたマイニング技術は、流行やトレンドを把握して新製品やサービス開発に活用したり、逆に株価情報の操作や製品に対するクレームなどネット上に広がる風説流布をいち早く監視するためにも重要な技術となってきている。

我々のグループでは、Web 情報をマイニングすることによってホットトピックのモニタリングを行うシステムについて検討を行ってきた。Web 情報のマイニング手法としては、(1)アクセスログや検索ログなどの利用状況を分析するもの、(2)ハイパーリンクの構造を分析するもの、(3)ページのコンテンツ自身を分析するもの、というように３つに分類される[1]が、我々は文献[2]において (2) の手法について報告した。その中で今後の課題として、リンクを張るという作業が比較的コスト高であるため、実際のブームの立ち上がりとずれてしまう事を述べた。

そこで、我々はブームの立ち上がりとのずれが少ないと思われる(1)の検索ログを使ったアプローチを現在進めているが、大量かつ雑多な検索ログの中からホットトピックの抽出対象となるログに絞り込むことは非常に困難である。本稿では、大量のログの中からホットトピックを抽出する上で重要となるサイトのフィルタリングを行うために、(時間、検索キーワード、飛び先 URL) という 3 つ組の検索ログデータを解析することによって、サイト種別(ニュース、大手 ISP、企業などのオフィシャルサイト)を自動的に判定し、話題の抽出対象を絞り込む手法を提案する。

2. 検索ログの特長

一般的に、検索サイトではユーザの検索行動の履歴である検索ログを保存しているが、これは人間の知的欲求行動の貴重な記録といえる。検索ログの研究については、検索クエリとそれが入力された時間のログを使ったものが主で、時間軸を使って検索語を同義語としてまとめることで情報のニーズの把握を高めるもの[3]やユーザの検索意図を推測するために検索パターンのモデル化を行ったもの[4]などがある。

本稿では、時間と検索クエリだけではなく、検索した結果の中からユーザが選択したページの URL を併せて {時間、検索クエリ、飛び先 URL} の三つ組みで記録し、ユーザ検索意図をより正確に捉えられるように工夫している。

WEB ページ中に書かれているアンカー文字列と URL の対では「このキーワードからはこのページを参照して欲しい」という情報公開者の意図が反映されているが、本稿の検索ログから抽出される検索キーワードと飛び先 URL の対では「このキーワードを使って知りたかったのはこのページである」という情報要求者の意図が反映されている。すなわち、「どの URL がどんなキーワードで検索され、実際に参照されたか」を表す本稿で扱う検索ログは、WEB ページやサイトの性格を判断する上で非常に有用な手

がかりとなる。

3. URL の参照パターンのモデル化

一般的にページの内容に着目すると、リンク集やニュース集のように雑多な内容が書かれたページと単一の内容について書かれたページに分けられる。それらのページを検索することを考えた場合、前者は多数のキーワードで検索され、後者は少数のキーワードで検索されることが予想される。これを踏まえると、検索クエリと URL の参照パターンは図 1 のような 3 つのパターンに大別される事が分かる。

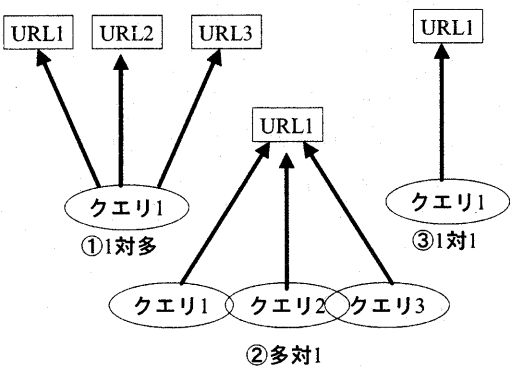

図 1 検索クエリと飛び先 URL の関係

① 1対多 ：1種類(または少種類)のクエリに対して多くの URL に飛ぶもの

② 多対 1 ：多種類のクエリに対して 1 種類(または少種類)の URL に飛ぶもの

③ 1対1 ：1種類(または少種類)のクエリに対して 1 種類(または少種類)の URL に飛ぶもの

本稿では、URL に対する異なりクエリ数に注目する。②のようなパターンは 1 つの URL を雑多なキーワードで参照しているので、複数の観点をもつページ、すなわちリンク集や多くのサービスを提供する ISP のページであることが予想される。また、③のようなパターンの URL ではユーザが常に 1 つの観点から参照しているページなので、定番・専門ページ(サイト)であることが予想される。

4. サイト種別の判定

上記で述べた URL の参照パターンを用いてサイトの種別判定を試みる。ここでサイトとは http://aaa.bbb.com/ や http://aaa.bbb.com/index.html のように URL 中の/で区切られる階層数が 1 または 2 である URL を指す。検索ログ中に含まれる各 URL を {URL へのアクセス数、URL を参照した異なりクエリ数、URL の階層数} の 3 つ組で集計して、以下のような仮説を立てる。なお、AND によって複数のキーワードで 1 つの検索クエリが構成されていてもそれは 1 つの検索クエリとみなし、またクエリ中に含まれるキーワードの表記揺れなどの対処は行っていない。

仮説 1 異なりクエリ数が少なくかつアクセス数が少ないサイトは、個人などでドメインを取得した小規模な専門サイトである。

仮説 2 異なりクエリ数が少なくかつアクセス数が比較的多いサイトは、企業や団体の定番的(オフィシャル)サイトである。

仮説 3　異なりクエリ数が多くかつアクセス数も多いサイトは、大手 ISP やニュース系またはリンク集サイトである。

仮説１に関しては、映画の前宣伝のために新規にドメインを取得して立ち上げたサイトや個人で立ち上げたばかり誹謗中傷サイトなどがこの範疇に入る可能性が高く、話題抽出という観点からは注目しておかなければならないサイト群である。仮説２に関しては、ユーザから固定的なイメージで見られる定番的なサイトであり話題抽出の対象となる可能性は低い。ただし、株価情報操作などである企業のトップサイトへのアクセスが急増することは十分考えられるためアクセス数の増減の監視はある程度必要である。仮説３に関しても雑多なイメージで見られる ISP や大手ニュース系サイトは話題抽出の対象から外してよい。

5. 実験

5.1 実験で用いた検索サイトの概要

本稿の実験に用いた検索サイトの利用概要を表１にまとめる。このサイトは一般コンシューマ向けたインターネット検索サイトである。

表 1 実験で用いた検索サイトの利用状況

総検索数	**約 50000 件/日**
新規検索数	**約 22500 件/日**
ユニーククエリ数	**約 7500 件/日**
AND 検索数	**約 2000 件/日**
ユニーク URL 数	**約 60000 件/日**
前日との一致度	**検索クエリ 10%** **飛び先 URL 10%**
3ヶ月前との一致度	**検索クエリ 5%** **飛び先 URL 10%**

総検索数は新規検索数に改ページによって生じる検索を加えたものである。新規検索との比が約1:2なので１検索当たり２ページ分の検索結果をユーザは参照していることがわかる。また、ユニーククエリ数と飛び先のユニーク URL 数の比から、１件の検索クエリによる検索結果の中からユーザは９件ものURLを参照していることがわかる。

5.2 実験と評価方法

4 章の仮説を検証するために実際の上記検索サイトの検索ログ(2001/9/23―2001/9/30)から、飛び先の URL 毎に｛アクセス数、異なりクエリ数、階層数｝を集計し、その結果を図２のグラフに示す。

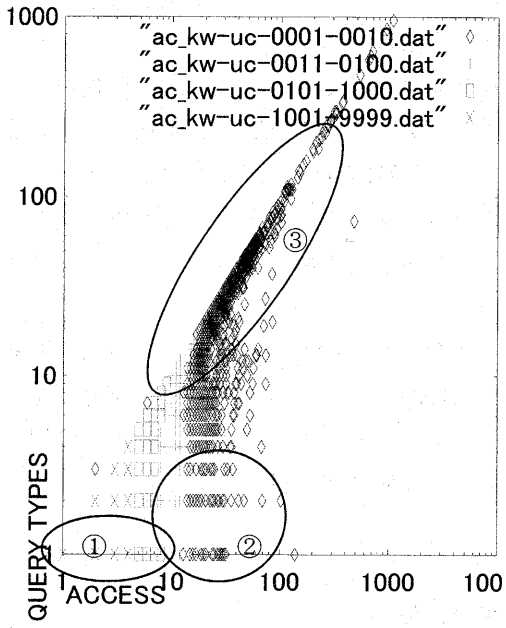

図 2 アクセス数ークエリ種別数グラフ

図 2 のグラフの xy 軸は共に LOGSCALE になっており、x 軸は URL に対するアクセス数を y 軸は検索クエリの種類数を表しており、実践で囲んだ①、②、③、の領域がそれぞれ、仮説 1−3 のサイト群に相当する。またプロット記号の◇、＋、□、×はプロットの重みを表しており、

◇　：頻度１－１０
＋　：頻度１１－１００
□　：頻度１０１－１０００
×　：頻度１００１以上

となっている。例えば、アクセス数が 1000 に近いところでは◇が多くプロットされているが、これはこのようなパターンの URL が１－１０（数として非常に少ない）ことを示している。

図 2 のグラフの①—③のそれぞれの部分から実際にいくつかの URL を取り出し、その中でも階層数が 1−2 のものについて、どのようなサイトであるかを人手で検証する。抽出したサイト例と検証結果を表 2 に示す。

表 2 評価結果

①のサイト例
アクセス数=1、　異なりクエリ数＝1

アクセス数	異なりクエリ数	URL
1	1	www.iwata-cc.co.jp
1	1	www.jadea.or.jp
1	1	www.igakueizou.co.jp
1	1	www.yasaiseikatsu.com/
1	1	www.chutoku.co.jp

②のサイト例
アクセス数＞5、　2＜異なりクエリ数＜5

アクセス数	異なりクエリ数	URL
11	1	www.sanrio.co.jp
13	3	www.shinseido.co.jp
9	3	www.meitetsu.co.jp
16	1	www.happo-en.com
17	1	www.jartic.or.jp

③のサイト例
アクセス数＞20、　異なりクエリ数＞20

アクセス数	異なりクエリ数	URL
1074	957	www.asahi-net.or.jp
939	854	www2s.biglobe.ne.jp
672	603	village.infoweb.ne.jp
644	576	www.geocities.co.jp
479	73	www.yahoo.co.jp/

5.3 考察

検索ログ中のユニークな URL の中で階層数が 1 のものは約 26000 であり、①, ②、③の条件に当てはまるものはそれぞれ① 16181、②183、③335 であり、②③の URL に関してはほぼ仮説を満たしていた。また、

過半数の URL が①に属するわけだがアクセス数が1であるため、検索結果の中から間違えて飛ぶというようなノイズ的な URL も含まれていると考えられその判定が難しいが、ある一定期間内でのアクセス数が増えるような場合のものが仮説1に含まれると考えられる。

6. まとめ

本稿では、インターネット上の情報から話題となるコンテンツをいち早く見つけるために、サイトを絞り込む手段として重要なサイト種別を検索ログの中の飛び先 URL と検索クエリを使って自動判定する手法を提案し、その有効性をネット上で検索サイトのログを使って適用し、その有効性を確認した。

7. 今後の課題

本稿では、階層数が1であるいわゆるトップサイトについてのサイト種別判定の指針を示したが、絶対多数である中間層のページについては本手法がどこまで適用できるかは確認できていない。中間層のページの中でも特に、注目を集めつつある個人ページは話題抽出の対象としては重要であり、これを自動的に判定する手法が望まれる。これを解決するアイデアとしては、サイト種別を判定したトップ URL の情報を用いて、個人ページと判断する候補ページを絞り込み、URL の各階層単位にアクセス数・異なりクエリ数を集計することで、個人ページのトップが特定できないかと考えている。

また、我々のグループの既存研究でもある WEB マイニングの手法と検索ログを使った本手法を併せることによってより精度の高い話題抽出を行う予定である。

参考文献

[1]Kosala,R. and Blockeel,H. : Web Mining Research: A Survey, ACM SIGKDD, Vol.2,No.1,pp.1-15(200).

[2] 西野,津田,牛,呉：Web マイニングによるホットトピックモニタリング，言語処理学会第8回年次大会(2002).

[3] 大久保,杉崎,井上,田中：WWW 検索ログに基づくトレンド情報の抽出,情報処理論文誌,Vol.39No.7,pp2250-2258(1998).

[4] 鈴木,山名：時間間隔を用いた検索履歴のモデル化,情処研報,DD032－014(2001).