4P-3

アクセス情報を用いたWEBサイトの流行判定とそのサイト推薦システムへの応用

吉村　武[†]　松本　和博[†]　内田 理[†]　中西 祥八郎[†]

Fashion judgment of WEB site using access information

Takeru Yoshimura[†] Kazuhiro Matumoto[†] Osamu Uchida[†] Shohachiro Nakanishi[†]

1. はじめに

近年インターネットの普及に伴い新たなる情報伝達メディアとしてWEBサイトの利用が急増してきた．それによりWEBサイトの数も爆発的に増加している．だがインターネット利用者にとっては無数に増えていくWEBサイトの中から必要な情報だけを抜粋し閲覧することは、現状では不可能に近い．そこで話題性の高いWEBサイトを抽出することができれば、ユーザーは自分が必要としている情報量の多いWEBサイトを閲覧する事が可能となる．

現在WEBサイトを閲覧する際、ユーザーは主に「YAHOO！」や「Google」などのポータルサイト上の検索システムを使用し、WEBサイトを閲覧している．ただしこのような既存の検索システムを利用した時には、利用者が必要としている情報を掲載しているWEBサイトを見つける事が困難である場合が多い．このような問題に対処するため、各ユーザーの特性を利用したWEBサイト推薦に関する研究などが行われている（例えば[3]）．

本研究では、サイトの閲覧者数の情報であるWEBサイトのアクセス情報を利用する事により、膨大なWEBサイトの中から話題性のあるWEBサイトを容易に閲覧することが可能となる方法を提案する．

2. WEBサイトの流行判定

2.1 現在の検索システム

従来の情報検索サービスを利用して目的の情報を持つWEBサイトを検索するためには、ある程度の知識と手間が必要とされる．なぜなら目的の情報を得るためには、関連する単語を入力しなければならず、さらに検索された複数のWEBサイトの中から必要な情報が存在するかを確認する必要も生じるからである．しかもWEBサイトの特徴として、どのような情報も掲載できるので、その情報の真偽をも確認しなければならない．このようにして個人での情報検索は情報の検索だけでなく、似たようなサイトの情報との検証を行いながら、どのサイトが一番自分の目的に合うのかを見極める必要がある．また、検索されてきた情報が大量であることも稀ではなく、目的の情報が後半にあった場合にそれを見つけることは困難である．そこで、これらの改善方法としてGoogle [1]では、PageRank [2]と呼ばれるWEBページのリンク構造を生かしたアルゴリズムにより上位に表示するWEBサイトを決定している．PageRankとは、リンクを支持投票とみなしその投票数により順位を決定するという考え方に基づいている．だが、これではサイトに目的の情報が詳細に掲載されていたとしてもリンクが無ければ上位に反映されることはなく、下位に位置付けられてしまうという問題がある．

2.2 流行のサイトの抽出

本研究は、アクセス数を利用することにより、話題性のある流行のWEBサイトを抽出する事を目的としている．

アクセス数はWEBサイトを閲覧すると増えるため、WEBサイトの人気を表す最も基本的なデータであると位置づけられる．例えばアクセス数が急増しているサイトは、現在利用者が必要としている情報が掲載されているWEBサイトであると考えられる．そこで、アクセス数を使い、話題性のある情報を持つ流行WEBサイトを抽出すれば、利用者が必要としているWEBサイトを閲覧することが容易になると思われる．

2.3 アクセス情報の解析

本研究では実際に運用されているWEBサイトのアクセス数を用いて話題性のあるWEBサイトを選別する方法を提案する．

WEBサイトの閲覧者数を表すアクセス数は様々な要因により常に変化している．その主な要因の一つとして、WEBサイトの内容が話題性の高い内容を持つ時に上昇する事が多い事が挙げられる．そこで数あるWEBサイトの中からアクセス数の上昇率の高い時期のWEBサイトのみを集める事により、話題性のあるWEBサイトを集めることが可能となると思われる．ただしアクセス数は日々の変動が激しく、一日ごとの変化率を求めても話題

[†]東海大学電子情報学部情報科学科
School of Information Technology and Electronics, Tokai University

性のある WEB サイトを集めることは不可能である．そこで WEB サイトの一定期間の平均と当日のアクセス数を比較する事で、その WEB サイトに話題性のある情報が掲載されているということを確認することが可能と考えた．

その他の問題として、WEB サイトに記載されている内容により目的の情報を必要としている人数に差が生じる事である．WEB サイトはアクセス数が少なくなるに従いアクセス数の変動が激しく、アクセス数が多くなるに従いアクセス数の変動が少ない傾向があり、上昇率がどれだけ大きくても、アクセス数が非常に少ない時には話題性がある WEB サイトであるとは言えない．この問題を解決するために、アクセス数と上昇率の両方を考慮したバランスの良い重みをかける必要がある．そこでアクセス数を尊重する重みに $\log_{10} H$ （H は当日のアクセス数）を使用し、変化率に掛ける事により全ての WEB サイトを同じ基準で比較することができる．

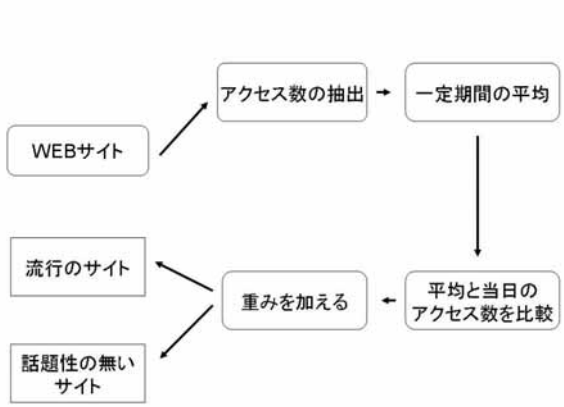

図１．流行 WEB サイト抽出手順

2.4 流行WEBサイト抽出アルゴリズム

以下に具体的な手順を説明する

(1)WEB サイトの当日のアクセス数 H を抽出しデータベースに格納する

(2)一定期間 n のアクセス数の平均 $E(H_n)$ を求める

(3)アクセス数 H を平均 $E(H_n)$ で割り変化率 P を求める

(4) P に重み $\log_{10}(H)$ をかけ流行判定基準 R を求める

$$R = P \times \log_{10}(H)$$

2.5 実験結果と考察

15 個の WEB サイトのアクセス数に対して一日ごとの R を求め、以下に R の分布を表す．

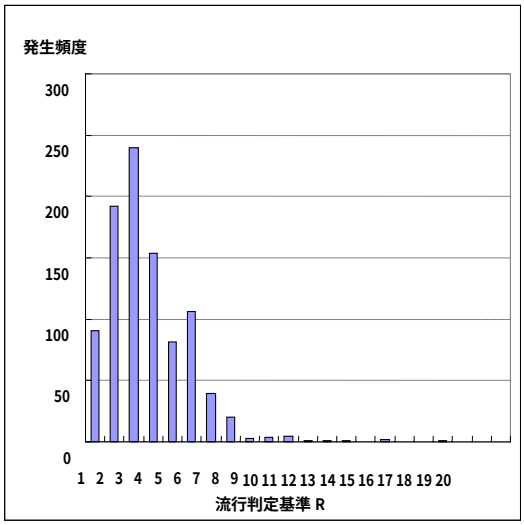

図 2. 流行判定基準 R の分布

図 2 より、流行判定基準 R の値が大きくなることは少ないと判断できる．流行判定基準 R が高い WEB サイトは長期的に見てアクセス数が上がっていることが確認できた．すなわちこの流行判定基準 R を使うことにより話題性のある WEB サイトを判断する事ができると考えられる．

3. まとめと展望

流行判定基準 R を求めることにより WEB サイトの長期的に上昇しているかを判別することができた．この事から無数に存在する WEB サイトの中から利用価値の高い情報を持つ WEB サイトの選別ができると言える．今後はアクセス情報を利用し明確に推薦する方法を検討して行きたい．

参考文献

[1]Google: http://www.google.com/

[2] L. Page et al.: The PageRank Citation Ranking: Bringing Order to the Web, http://www-db.stanford.edu/~backrub/pageranksub.ps

[3]松本和博，與良光一郎，内田理，中西祥八郎：”パーソナル化を利用した WEB サイト推薦システム”，第 67 回情報処理学会，3U-7(2005/03)