時系列情報を考慮したアクセスログ解析*

4S－9

畑島 隆†　元田敏浩‡
日本電信電話株式会社 ソフトウェア研究所
takashi@canary.sl.cae.ntt.co.jp　motoda@canary.sl.cae.ntt.co.jp

1. はじめに

WWWコンテンツへのアクセス数の計測手法では、サーバのアクセスログを解析する手法がある。また、アクセス履歴からユーザモデルを構築[1]し、インターネット上での one to one marketing 等に適用する試みがある。しかしこれらの手法では、いずれも一定期間のアクセスの総量を算出し、期間中のアクセスを平等に扱うため、アクセス履歴情報が損なわれている。

本稿ではアクセス履歴の時系列を考慮した評価指標「関心度」を提案し、適用領域を明らかにする。

2. 関心度の定義

本稿で提案する関心度は、WWWアクセスログの時系列情報を考慮[2]し、コンテンツへのアクセス履歴を時間関数により得点化したものであり、以下の項目で定義される。

①コンテンツへのアクセス間隔

②コンテンツに対する直前のアクセス時点の関心度

アクセスが集中すると関心度は高くなり

以下に関心度の評価式の定義を述べる。

2.1 評価式の定義

関心度ではアクセスによる増加と時間の経過による減衰が同時に作用する。関心度の評価式の要求条件を以下の時間関数により定義する。

直前のアクセスからの時間間隔 Δt について、アクセスによる関心度の増加あらわす関数を $g(\Delta t)$ とする。また、時間経過による関心度の減衰を、直前のアクセス時点の関心度との比率によってあらわす関数を $h(\Delta t)$ とする。このときアクセスの発生していない任意の時刻 $t(t_n < t < t_{n+1})$ の関心度 $F(t)$ を次式で定義する。

$$F(t) = h(\Delta t)F_n \qquad \text{ただし、} \Delta t = t - t_{n-1} \tag{1}$$

$$F_n = g(\Delta t) + h(\Delta t)F_{n-1} \quad \text{ただし、} \Delta t = t_n - t_{n-1} \tag{2}$$

ただし F_n はコンテンツに対する n 回目のアクセス(時刻 t_n)時点での関心度である。

また、各関数の要求条件として以下の式を定義する。

・関心度刺激関数 $g(t)$

$$g(t) > 0, \quad \text{ただし } g(0) = \max(g(t)) \text{（定数）}$$
$$g'(t) \le 0 \tag{3}$$

・関心度減衰関数 $h(t)$

$$0 \le h(t) \le 1, \text{ ただし } h(0) = 1$$
$$h'(t) < 0, h''(t) \ge 0 \tag{4}$$

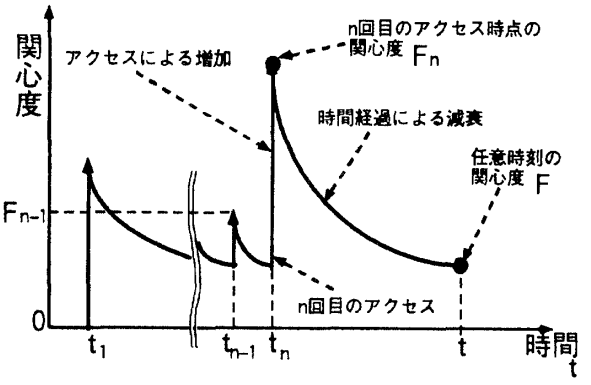

図1:関心度の変化

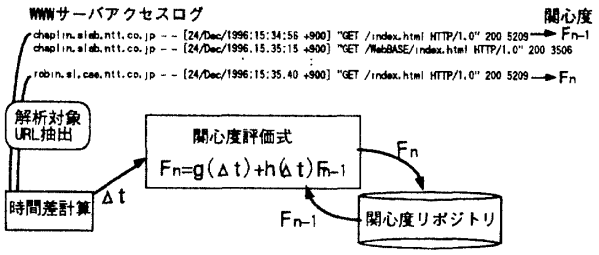

図2:関心度評価システムの構成図

式1で提案した関心度の変化の様子を図1に示す。また、解析対象をコンテンツのURLで指定する場合の関心度評価システムのの構成は、図2のようになる。以下に本節の定義にしたがって、関心度刺激関数と関心度減衰関数をモデル化した関心度評価式の例を述べる。実際のWWWユーザのアクセス動向を調査し、アクセスによる関心の変化の様子を調査した。

2.2 アクセスによる関心度増加

ユーザによるコンテンツへのアクセスにより、関心度が高められる。変化量はアクセスの時間間隔や、それまでの関心度の値によって決定される。しかし、今回はアクセスによる刺激を時間間隔によらず一定とし、関心度刺激関数を次式で定義する。

$$g(\Delta t) = const. \tag{5}$$

2.3 時間経過による減衰

前述の定義により、あるコンテンツに対する関心度は、そのコンテンツに対してアクセスが存在しない限り減衰する。関心度減衰関数をモデル化するために、WWWサーバのアクセスログから実際のWWWユーザのアクセス動向の調査を行った。調査対象としてNTT DIRECTORY[1] のサーチエンジン InfoBee[2] への検索キーワード投入件数を用いた。キーワードが現わすイベント終了後同じイベントが発生しない期

---

* An Analyzing of the Access_logs Regarding its Time Sequence Data.

† Takashi Hatashima, NTT Software Laboratories.

‡ Toshihiro Motoda, NTT Software Laboratories.

[1] NTT DIRECROTY, http://navi.ntt.co.jp/

[2] InfoBee, http://navi.ntt.co.jp/j/info/infobee.html

間において、時間経過による検索数の減少が顕著に現れるキーワード「選挙」[3] に着目した。選挙終了後の検索数の時間経過による減衰を、関心度の減衰の一例として関心度減衰関数に用いる。

また、昼間・夜間といった時間帯による検索数の変動を正規化するために以下の式であらわされるアクセスシェアを各時間帯で求めた。

$$アクセスシェア = \frac{当該時間帯の対象語検索数}{当該時間帯の総検索数} \quad (6)$$

関心度の減衰を①情報の利用度の低下[3]②情報の老化[4]③記憶の忘却[5]の観点から考察した結果、知見において減衰モデルは指数関数により定義されているため、関心度減衰関数に指数関数を用いた。

時間経過による自然な減衰傾向を示すと考えられる期間のアクセスシェアの減衰を近似して求められた時間関数を関心度減衰関数とした。投票終了後、開票情報に対する検索要求が発生し、20 日 19 時台にピークとなった。翌日起床後に情報検索するユーザを除外するため、アクセスシェアが増加する直前の21日午前4時までを選択した。図3にこの期間のアクセスシェアの実数と関心度減衰関数を示すように、ピークとなった時刻（$t = 0$）以降の時間経過を Δt [hour]として、前述した定義により$h(0) = 1$を切片とすると、関心度減衰関数は次式のように導かれる。

$$h(\Delta t) = \exp(-0.2765\Delta t) \quad (7)$$

今回は関心度刺激関数 $g(\Delta t)$ を定数とし、1回のアクセスにつき1点の関心度値が加算されるとしたため $g(\Delta t) = 1$ となるので、n 回目のアクセス以降の任意の時刻 $t(t_n \le t < t_{n+1})$ における関心度 $F(t)$ は次式で定義される。

$$F(t) = F_n \exp(-0.2765\Delta t) \quad ただし \Delta t = t - t_{n-1} \quad (8)$$

$$F_n = 1 + F_{n-1}\exp(-0.2765\Delta t) \quad ただし、\Delta t = t_n - t_{n-1} \quad (9)$$

なお、F_n はあるコンテンツに対する n 回目のアクセス（時刻 t_n）時点における関心度であり、Δt の単位は[hour]である。

3.関心度の適用領域

関心度の以下の事例へ適用を検討している。

・キャッシュ管理

NTT DIRECTORY [1]のようなWWWと DB の連携により検索サービスを提供するシステムでは、検索結果が逐次キャッシングされる。キャッシュ管理に関心度を適用し、関心度の高い検索結果だけをキャッシングする。これにより最近多く検索された内容だけがキャッシングされるため、キャッシュ利用効率の向上が見込まれ、より多くの検索要求に対して高速応答が可能になる。

・コンテンツランキングの提供

関心度の高いコンテンツをランキング表示することにより、人気コンテンツを表示するサービスを提供する。

・コンテンツ鮮度管理

関心度の低いコンテンツに対してしきい値を設定することにより更新や廃棄の基準とする。

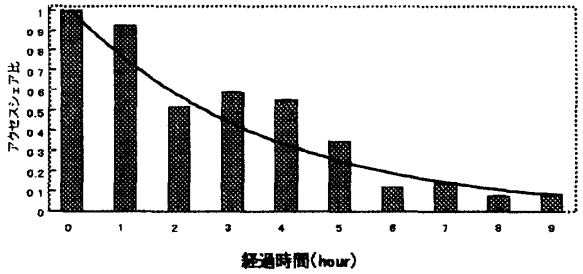

図 3:アクセスシェアのピーク以降の時間変化と指数近似のグラフ

なお、関心度評価値の算出にはアクセス履歴の時間情報のみを用いているため、WWWサーバに限らずアクセスに対してタイムスタンプを発行する機能を持つサーバのコンテンツ管理に適用可能である。

また、関心度は同一サーバ上のコンテンツ間の相対評価の指標であるため、同一条件でのアクセスログ解析が可能なWWWサーバ内のコンテンツ管理に有効な指標であるといえる。しかし、複数の WWW サーバについての解析では、I／PRO[4] のように他サーバのアクセスログを関心度解析サーバに転送し解析することにより、公正な評価手段として適用可能となる。

4. 課題

関心度減衰関数の関数のモデル化に用いた検索キーワード「選挙」によって、

①短期間にアクセスが集中

②単発のイベントであり他のイベントとの関連が薄い

という性質のはっきりした事象に対する情報要求をモデル化したが、実際の関心度の減衰に近づけるためにはより多くの事象について調査する必要がある。

また関心度刺激関数を定数にしているため、アクセスの集中により関心度が無限大に発散してしまうことが予想される。しかし実際にはあまり多く同じ事象に触れると「飽き」がくるように、関心度についても増加の上限があると考えられるため、これを考慮した関心度の成長モデルを定義したうえで関心度刺激関数を定める必要がある。

5. まとめ

本稿ではWWWサーバへのアクセスの時系列を考慮したアクセス動向調査指標として関心度を提案した。実際のアクセスログにおけるアクセス件数の減少傾向から関心度減衰関数を導いた。また現在検討中の実システムへの適用例を示した。今後は多くの事例に検討しモデル化を進めるとともに、実システム上での評価を行う予定である。

参考文献

[1] 三浦,高橋,島."個人適用型WWWのためのユーザモデル構築", 情報処理学会, Interaction '97,Feb,1997

[2] 畑島,元田."WWWアクセスログの有効な解析法について", 第53回情処全大,分冊 3 pp.217-218,1996

[3] Cole. P. F, "Journal usage versus age of journal", Journal of Documentation, 19, pp.1-11

[4] Griffith, B. C. et al. ,"Aging of scientific literature : a citation analysis", Journal of Documaentation,35,1980, pp179-196.

[5] 古川,"寿命の数理",行動計量学シリーズ 13,朝倉書店, 1996,pp.191-196

[3] 第 41 回総選挙, 告示 1996.10.8, 投票時間 1996.10.20,7:00-18:00 (即日開票)

[4] I/PRO, http://www ipro.com/