PAPER *Special Issue on Knowledge-Based Software Engineering*

An “Interest” Index for WWW Servers and CyberRanking

Takashi HATASHIMA[†], **Toshihiro MOTODA**[†], ***Nonmembers,***
and **Shuichiro YAMAMOTO**[†], ***Member***

SUMMARY We describe an index for estimating the level of interest in Web pages. This “time-based interest” (TBI) index combinates an equation reflecting page accesses and an equation reflecting the decrease in interest over time. These equations work simultaneously by using a parameter that is based on the time since the last access. We experimentally estimated the decrease ratio of the TBI index and evaluated the characteristics of the TBI equation. We found that the index follows Zipf’s distribution, indicating that reflects the change in popularity. We also introduce an access-log analysis system called CyberRanking that includes TBI analysis. CyberRanking analyzes the access logs of Web servers and presents the results in 2-D or 3-D graph on a Web browser.

key words: *access log analysis, time based interest, World Wide Web, visualization*

1. Introduction

As commercial use of the Internet has grown, so have the demands for determining the value of World Wide Web sites and the popularity of Web pages. Existing indices such as page view (the number of the times a Web page is viewed), Visit (the muber if time a page is viewed distinguished from from the next by a ‘timeout’.), and “the access path of a user” can be determined from the access logs of Web servers. Analyzing of these indices is used for one-to-one marketing or for personalizing Web pages. However, the calculation of these indices does not take into account the sequence of time-stamps in the access logs, so this analysis is not sufficient for determing the level of interest users have in the contents of the site.

Current services such as searching (e.g. Yahoo! [1]), site ranking (e.g. 100hot.com [2]), and personalization (e.g. MyYahoo! [3]) also do not use the sequence of time-stamp in their access logs. It is thus difficult to determine the degree if user interest in their pages.

We developed an index for Web pages that makes use of the sequence of time-stamps in the access logs.[4]. This “time-based interest” (TBI) index is defined as the sum of two equations, each with a variable representing when the page was last accessed. We have now applied this index to a directory service and evaluated its usefulness.

We have also developed the CyberRanking directs users to pages of potential interest, enables real-time analysis by retrieving access logs stored in databases, and allows easily customized analyses. The CyberRanking visualizaton system was designed to analyze access logs based on the TBI index.

2. Time-Based Interest Index

The index reflects the popularity of the Web is time-based interest (TBI). The index for a page goes up every time the Web page is accessed. It goes down over time as the page contents become outdated.

2.1 Calculation of TBI

The TBI index F_n is the sum of the increase in the number of accesses and the decrease in accesses over time:

$$F_n = g(\delta t) + h(\delta t)F_{n-1}, \qquad (1)$$

where n is the number of accesses to a page, t is the time, $g(\delta t)$ is the increase value and $h(\delta t)$ is reduction ratio, and δt is the time span between the $n-1$-th and n-th accesses ($\delta t = t_n - t_{n-1}$).

For a period in which no user accesses the page, $F(t)$ at time t is given by

$$F(t) = h(\delta t)F_n, \qquad (2)$$

where $\delta t = t - t_{n-1}$, t is the time between accesses ($t_n < t < t_{n+1}$), and δt is the time since the page was last accessed ($\delta t = t_n - t_{n-1}$). Figure 1 shows an example of how a change in interest over time is reflected in the TBI index shown as $F(t)$ and as its discrete form F_n. Figure 2 shows how the TBI index is calculated. The URLs of the individual pages to be audited are extracted from the server’s access log. The TBI index is calculated as a function of the elapsed time, and the current value stored in a repository.

2.2 Increase Equation

The increase equation for the TBI index shows how it increases every time a user accesses the page. This equation is used only when a page for which a TBI

Manuscript received August 20, 1999.
Manuscript revised November 5, 1999.

[†]The authors are with NTT Information Sharing Platform Laboratories, Musashino-shi, 180–8585 Japan.

index is calculated.

We defined two requirements for increase equation $g(t)$. 1) Do not decrease the index value. 2) When the time between accesses is zero (i.e., simultaneous access), increase the index by the maximum amount.

$$
\begin{aligned}
&g(t) > 0\\
&g(0) = max(g(t))(const.)\\
&g'(t) \leq 0
\end{aligned}
\tag{3}
$$

The value calculated depends on time parameter δt and the value calculated when the page was last accessed. In this paper, we set $g(t)$ as a constant.

$$g(\delta t) = const. \tag{4}$$

In future work we will calculate it.

2.3 Decrease Equation

2.3.1 Overview

The decrease equation is used to calculate how much the interest level in a page has declined. We defined three requirements for decrease function $h(t)$. 1) It should reduce index monotonouslly. 2) The reduction ratio must be positive. 3) Do not decrease the index value when users access the page simultaneously.

$$
\begin{aligned}
&0 \leq h(t) \leq 1\\
&h(0) = 1\\
&h'(t) \leq 0\\
&h''(t) \geq 0
\end{aligned}
\tag{5}
$$

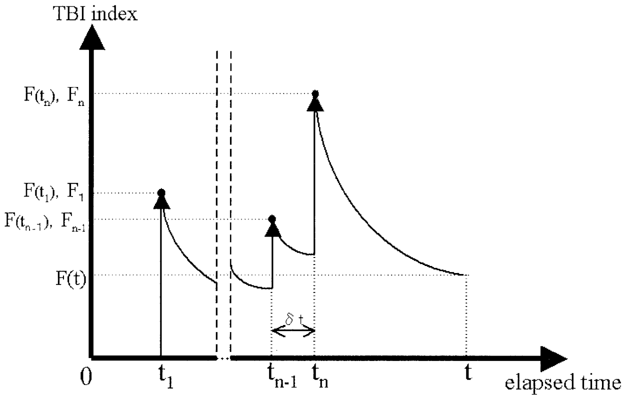


Fig. 1 Example of how change in interest over time is reflected in the TBI index.

This equation reflects how the level of interest in certain information declines due to decreasing usage, obsolescence, and decreasing interest.

2.3.2 Decreasing Usage of Information

Because users access pages to get information, their contents should be kept up to date or deleted when they become out of date. Libraries have a limited storage capacity, so deciding which books to keep and which to dispose of to make room for new ones is always a problem. Cole [6] showed that the usage of periodicals in a library decreases exponentially with time:

$$R(x) = R(N)e^{-Lx}, \tag{6}$$

where $R(x)$ is the number of requests for magazines published x years ago, $R(N)$ is the total number of requests for all magazines and N and L are constants.

2.3.3 Obsolescence

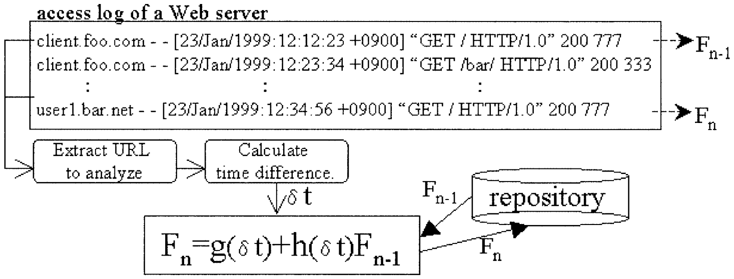
Information becomes outdated, so it has a certain life span. Griffith et al. [7] investigated the range of dates of documents cited in theses and found that the number decreased exponentially over time. We can thus say that the interest level in information goes down as the information becomes obsolete over time.

2.3.4 Decreasing Interest

Rumors, trends, and current events can cause a rapid increase in interest levels in the short term. However, people soon lose interest. Cognitive science research has shown that memory decays through disuse. The "retention ratio of memorization" is exponentially related to elapsed time, t [8]:

$$R(t) = R(\infty) + \{R(t) - R(\infty)\}e^{-\alpha t}, \tag{7}$$

where $R(\infty)$ is the constant as $t \rightarrow \infty$ in $R(t)$, and α is the reduction ratio.



```
client.foo.com - - [23/Jan/1999:12:12:23 +0900] "GET / HTTP/1.0" 200 777
client.foo.com - - [23/Jan/1999:12:23:34 +0900] "GET /bar/ HTTP/1.0" 200 333
      :                       :                              :
user1.bar.net - - [23/Jan/1999:12:34:56 +0900] "GET / HTTP/1.0" 200 777
```

Fig. 2 Calculation of TBI index.

2.3.5 Reduction Ratio Equation

The decrease equaton can thus be shown as an exponential equation of elapsed time. It shows the reduction ratio of the TBI.

Because we set $h(0) = 1$ (Eq. (5)), reduction ratio function is:

$$h(\delta t) = e^{-\alpha \delta t} \tag{8}$$

where α is the reduction ratio, and δt is the time since the last access.

2.4 Estimation of Reduction Ratio

We have experimentally estimated the reduction ratio by using the logs of the full-text search engine in NTT DIRECTORY [5], a popular portal site in Japan. The keywords entered by users were assumed to reflect the topics of interest. The total number of keywords entered for a topics was assumed to reflect the level of interested for all users. The reduction ratio was assumed to be the reduction in the ratio of queries for a event against all queries.

The keywords related to a topics were identified using a thesaurus. The set of Japanese words related to “election” was used to estimate the reduction in interest. Elections are held periodically and capture people’s attention, but the quickly lose it after the vote are counted, so they are a suitable example for estimating the reduction.

Because the interest in an election usually falls when the results are announced, the reduction ratio can be estimated as the reduction ratio of keywords submitted to retrieve Web pages about the election. The number of keywords submitted depends on the measurement period. To exclude this effect, we based our estimation on the access-share, that is, the ratio of target keywords to total keywords within a certain period (one hour).

$$Access - share = \frac{(Number\ of\ target\ keywords)}{(Number\ of\ all\ keywords)} \tag{9}$$

Figure 3 shows the change in the access-share for an election over time following the election. Given Eq. (5), the series of plot points was normalized so that the peak is at $h(0) = 1$. The reduction ratio equation is:

$$h(\delta t) = e^{-0.2765 \delta t}. \tag{10}$$

2.5 TBI Equation

Based on the discussion in Sect. 2.2, we set the increase function so that the TBI index increases by one point for each access, i.e., at a as constant $(g(\delta t) = 1)$. The TBI index at any time t after the n-th (n is a natural number) access to a page:

$$F(t) = F_n e^{-0.2765 \delta t}, \tag{11}$$

where $\delta t = t - t_n$. The TBI index at the time of the n-th (n is natural number) access to the page is:

$$F_n = 1 + F_{n-1} e^{-0.2765 \delta t}, \tag{12}$$

where $\delta t = t_n - t_{n-1}$.

2.6 Experimental Evaluation

We tested whether the TBI index can substitute for page view, the metric commonly used to rank Web sites by their number of accesses. A study on the popularity of Web documents [9], using page view as the metric, showed that the number of accesses to documents related to their overall rank in popularity followed Zipf’s law [10].

We investigated whether the TBI index also follows Zipf’s law. We used the access logs of NTT DIRECTORY from March 23 to June 9 1999. The directory included about 230,000 links at that time. We computed the TBI and page views indices of each URL. The results are shown as a dual-logarithmic diagram in Fig. 4. Each data point represents one page, with

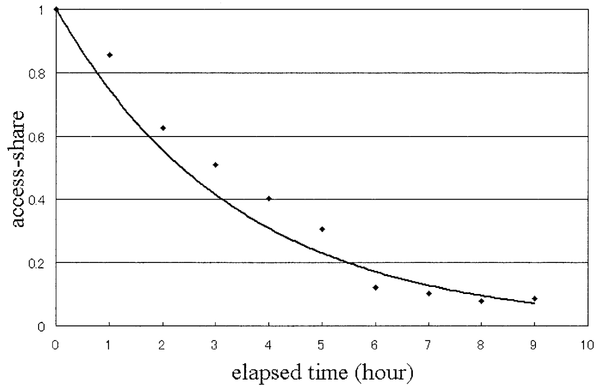


Fig. 3 Change access-share over time following an election.

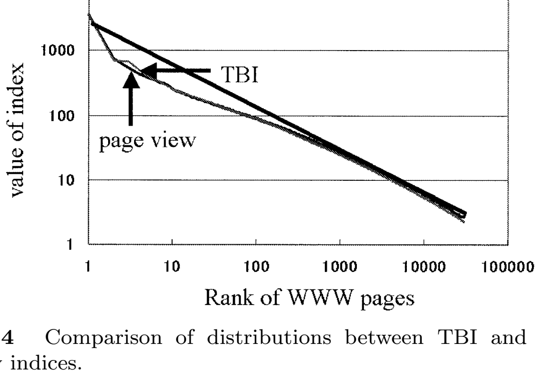
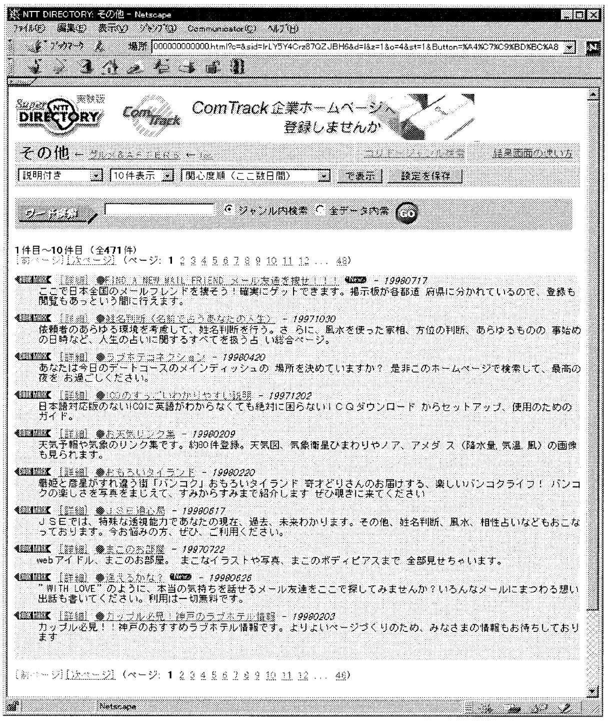


Fig. 4 Comparison of distributions between TBI and Page View indices.


NTT DIRECTORY: その他 - Netscape
ファイル(F) 編集(E) 表示(V) ジャンプ(G) Communicator(C) ヘルプ(H)
ブックマーク 場所 000000000000.html?c=&sid=IrLY5Y4Crz87QZJBH6&d=l&z=1&o=4&st=1&Button=%A4%C7%C9%BD%BC%A8
Super NTT DIRECTORY 実験版
ComTrack
ComTrack 企業ホームページへ 登録しませんか
その他 ← グルメ＆ＡＦＴＥＲ５ ← Top
結果画面の使い方
説明付き
10件表示
関心度順（ここ数日間）
で表示
設定を保存
ワード検索
◉ ジャンル内検索 ○ 全データ内索
GO
Netscape

1件目〜10件目（全471件）
[前ページ] [次ページ]（ページ: 1 2 3 4 5 6 7 8 9 10 11 12 ... 48)

BOOK MARK [詳細] ●FIND A NEW MAIL FRIEND メール友達を探せ！！！ NEW - 19980717
ここで日本全国のメールフレンドを捜そう！確実にゲットできます。掲示板が各都道 府県に分かれているので、登録も閲覧もあっという間に行えます。

BOOK MARK [詳細] ●姓名判断（名前で占うあなたの人生） - 19971030
依頼者のあらゆる環境を考慮して、姓名判断を行う。さ らに、風水を使った家相、方位の判断、あらゆるものの 事始めの日時など、人生の占いに関するすべてを扱う占 い総合ページ。

BOOK MARK [詳細] ●ラブホテコネクション - 19980420
あなたは今日のデートコースのメインディッシュの 場所を決めていますか？ 是非このホームページで検索して、最高の夜を お過ごしください。

BOOK MARK [詳細] ●ICQのすっごいわかりやすい説明 - 19971202
日本語対応版のないICQに英語がわからなくても絶対に困らないＩＣＱダウンロード からセットアップ、使用のためのガイド。

BOOK MARK [詳細] ●お天気リンク集 - 19980209
天気予報や気象のリンク集です。約80件登録。天気図、気象衛星ひまわりやノア、アメダ ス（降水量, 気温, 風）の画像も見られます。

BOOK MARK [詳細] ●おもろいタイランド - 19980220
織姫と彦星がすれ違う街「バンコク」おもろいタイランド 奇才どりさんのお届けする、楽しいバンコクライフ！ バンコクの楽しさを写真をまじえて、すみからすみまで紹介します ぜひ覗きに来てください

BOOK MARK [詳細] ●ＪＳＥ透心局 - 19980617
ＪＳＥでは、特殊な透視能力であなたの現在、過去、未来わかります。その他、姓名判断、風水、相性占いなどもおこなっております。今お悩みの方、ぜひ、ご利用ください。

BOOK MARK [詳細] ●まこのお部屋 - 19970722
webアイドル、まこのお部屋。 まこなイラストや写真、まこのボディピアスまで 全部見せちゃいます。

BOOK MARK [詳細] ●逢えるかな？ NEW - 19980626
”WITH LOVE”のように、本当の気持ちを話せるメール友達をここで探してみませんか？いろんなメールにまつわる想い出話も書いてください。利用は一切無料です。

BOOK MARK [詳細] ●カップル必見！神戸のラブホテル情報 - 19980203
カップル必見！！神戸のおすすめラブホテル情報です。よりよいページづくりのため、みなさまの情報もお待ちしております

[前ページ] [次ページ]（ページ: 1 2 3 4 5 6 7 8 9 10 11 12 ... 48)

Fig. 5 Application of the TBI index to NTT DIRECTORY.

the x axis showing the pages sorted according to their popularity by the indices. The x-y plots are very similar, and their equations shows that both distributions follow Zipf's law.

TBI index:

$$y = 7927.8x^{-0.6857}$$
$$R^2 = 0.9943 \tag{13}$$

Page view:

$$y = 7571.1x^{-0.6691}$$
$$R^2 = 0.9923 \tag{14}$$

In each equation, x is the rank ordered by each index, y is the value of the index. The TBI index can thus be used as an index for rating the popularity of Web pages.

2.7 Application of the TBI

We applied the TBI index to NTT DIRECTORY. Three different parameters of the decrease ratio were used to determine the variation in ranking. The hyperlinks in Fig. 5 were sorted by the TBI index computed with Eq. (12). This site has five sorting functions: by date, by title, by the TBI index, by URL, and by number of bookmarks stored in this site. As shown in Fig. 6, the TBI index was used the most for sorting.

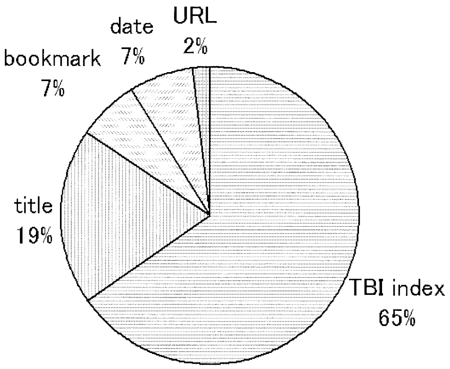


Fig. 6 Usage ratio of five sorting functions in NTT DIRECTORY.

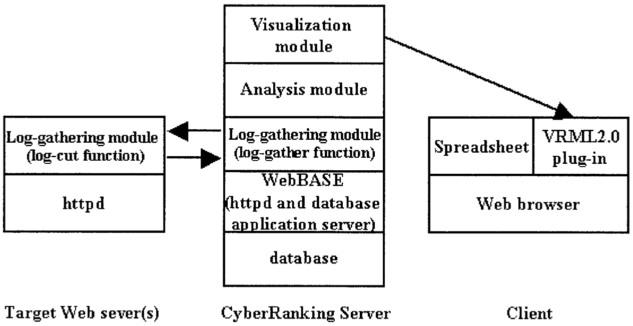


Fig. 7 Components of CyberRanking.

3. CyberRanking Access-Log Analysis System

3.1 Overview

The CyberRanking access-log analysis system we developed uses the TBI index and statistical analysis. As shown in Fig. 7 it has three modules. A log-gathering module collects the access logs of the target Web servers. An analysis module in the CyberRanking server analyzes the indices. A visualization module in the CyberRanking prepares the results for display on the user's browser.

3.2 Log-Gathering Module

This module has two functions. A log-cut function is installed in the Web servers to analyze their access logs. It also sends the access logs of the servers to a directory where the CyberRanking server can obtain them by HTTP requests. This is necessary because access logs are usually generated in a directory that cannot be accessed by an HTTP requests.

A log-gather function is installed in the CyberRanking server to gathers the access logs sent by the log-cut function. It collects the site map and page titles of the target servers by sending "robots" to the servers. The data obtained is processed by an application server

Table 1 Indices analyzed by CyberRanking.

Index	Units of analyzing the index	Classifications of units
page view	all accesses	1,2,3,4,5
	grouped by user domain	1,2,3,4,6
Visit	all accesses	1,2,3,4,5
	grouped by user domain	1,2,3,4,6
TBI	all accesses	1,2,3,4,5
	grouped by user domain	1,2,3,4,6
pages accessed within 1 Visit	all accesses	5,6

1: all, 2: weekly, 3: daily, 4: hourly, 5: by domain, 6: by Web pages

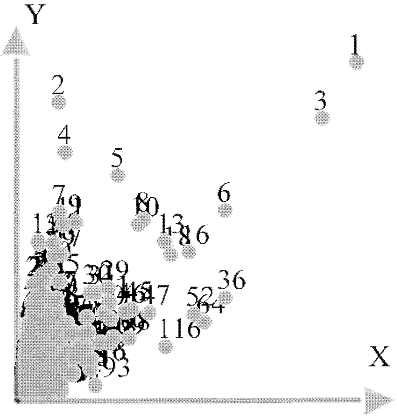


Fig. 8 Example of displayed result (2D).

(WebBASE [11], [12]) and stored in a database.

3.3 Analysis Module

The indices analyzed by the analysis module are listed in Table 1. They are classified according to time, such as the time of day or day of the week.

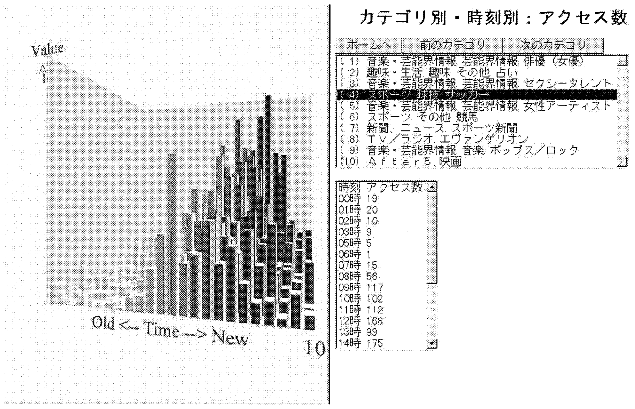
3.4 Visualization Module

This module prepares the analyzed data for display by a Web browser. Figures 8 and 9 show examples of displayed results. The user interface is written in JavaScript, the graphs are produced using VRML. This module can produce both 2-D and 3-D graphs.

4. Discussion

4.1 TBI Index

Related work on access-log analysis falls into two groups, 1) keyword-based interest, 2) user-based interest. For 1), Nomiyama et al. [14] proposed a ranking method that uses an index based on the weight of key words in a document and the frequency of access to them. However, the weight given to contents by the function at time is not considered. For 2), Shardanand and Maes [13] evaluated methods of forecasting the scores that individual users would assign to a contents based on the scores already assigned by other users. Kobayashi et al. [15] proposed a recommendation technique based on information the user selected. However, the neither change in interest over time was considered.



カテゴリ別・時刻別：アクセス数

(1) 音楽・芸能界情報 芸能界情報 俳優（女優）
(2) 趣味・生活 趣味 その他 占い
(3) 音楽・芸能界情報 芸能界情報 セクシータレント
(4) スポーツ 球技 サッカー
(5) 音楽・芸能界情報 芸能界情報 女性アーティスト
(6) スポーツ その他 競馬
(7) 新聞、ニュース スポーツ新聞
(8) TV／ラジオ エヴァンゲリオン
(9) 音楽・芸能界情報 音楽 ポップス／ロック
(10) After5 映画

時刻	アクセス数
00時	19
01時	20
02時	10
03時	9
05時	5
06時	1
07時	15
08時	56
09時	117
10時	102
11時	112
12時	168
13時	99
14時	175

Fig. 9 Example of displayed result (3D).

The only parameters needed to calculating the TBI index is the time since the last access and the index at that time. Because it is easy to calculate, it is superior to the other methods in terms of scalability.

The TBI index equation defined in this paper is experimental; there are several problems still to be studied. 1) In estimating the reduction ratio (Sect. 2.4), we used concentrated access in a short period. A more generalized reduction ratio is required. 2) The increase equation was set to a constant in this paper; it may also be an function of time; because a concentrated number of accesses may reduce the level of interest. 3) We did not consider the effect of revising or renewing the page contents on the TBI index. 4) Selecting suitable parameters to represent the characteristics of the users and pages is difficult and requires further study.

4.2 CyberRanking

CyberRanking presents a visualization of how Web pages are used, which can be useful for navigation. CyberRanking can work with OLAP (online analytical processing) systems, because it uses WebBASE, making it easy to retrieve their databases and handle them.

5. Conclusion

The time-based interest (TBI) index was designed to reflect the popularity of Web pages. The equation and functionality are also discussed. Moreover, We estimated the reduction ratio of the TBI as an exponential equation of the time span from the last access to the page. Experimental evaluation of the TBI index showed that is obeys Zipf's law, as does the number of accesses to pages. The CyberRanking visualizaton system was designed to analyze access logs based on the TBI index.

References

[1] Yahoo!, http://www.yahoo.com

[2] 100hot.com, http://www.100hot.com

[3] My Yahoo!, http://my.yahoo.com

[4] T. Hatashima and T. Motoda, "An analysis of the access_logs regarding its time sequence data," The 54th Natl. Conv. Rec. IPSJ Japan, 4S-9, Chiba Japan, 1997.

[5] NTT DIRECTORY, http://navi.ocn.ne.jp

[6] P.F. Cole, "Journal usage versus age of journal," J. Documentation, vol.19, pp.1–11, 1963.

[7] B.C. Griffith, P. Servi, A. Anker, and M.C. Drott, "Aging of scientific literature: A citation analysis," J. Documentation, vol.35, pp.179–196, 1979.

[8] T. Furukawa, "Jumyou no suuri," Behaviormetric Series 13, pp.191–196, Asakura Shoten, Japan, 1996.

[9] C.R. Cunha, A. Bestavros, and M.E. Crovella, "Characteristics of WWW client-based traces," Technical Report TR-95-010, Boston University Computer Science Department, June 1995.

[10] G.K. Zipf, Human behavior and the principle of least-effort, Addison-Wesley, Cambridge, MA, 1949.

[11] S. Yamamoto, R. Kawasaki, T. Motoda, and K. Tokumaru, "Internet/Intranet application development system WebBASE and its evaluation," IEICE Trans. Inf. & Syst., vol.E81-D, no.12, pp.1450–1457, Dec. 1998.

[12] WebBASE, http://webbase.ntts.co.jp/

[13] U. Shardanand and P. Maes, "Social information filtering: Algorithms for automating "Word of Nouth"," Proc. ACM Conference on Human Factors in Computing Systems (CHI'95), pp.210–217, 1995.

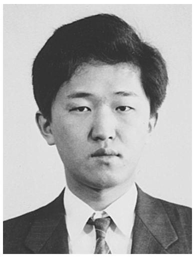
[14] H. Nomiyama, S. Konya, H. Watanabe, K. Kushima, and T. Tsutsumi, "Personalized information navigator: The hierarchical memory model for learning users' interests and its application to collaborative filtering," 42-8, IPSJ SIGFI Notes, pp.49–56, July 1996.

[15] K. Kobayashi, Y. Sumi, and K. Mase, "Information presentation based on individual user interests," Proc. IEEE Second International Conference on Knowledge-Based Intelligent Electronic Systems, pp.375–383, Adelaide, April 1998.

[16] C. Chen, "Structuring and visualizing the WWW by generalized similarity analysis," Proc. Eighth ACM Conference on Hypertext: Hypertext 97, pp.177–186, ACM Press, Southampton, U.K., 1997.

[17] S. Mukherjea and J. Foley, "Visualizing the World-Wide Web with the navigational view builder," http://www.igd.fhg.de/www/www95/papers/44/mukh/mukh.html, Special Issue of the journal COMPUTER NETWORKS AND ISDN SYSTEMS, vol.27-6, Germany, April 1995.

Takashi Hatashima is a stuff of NTT Information Sharing Platform Laboratories. His research interests includes distributed information systems and end user computing. Mr. Hatashima received a B.S. in information engineering and M.S. in geotechnical and environmental engineering from Nagoya University in 1993 and 1995. He is a member of ACM and IPSJ.



Toshihiro Motoda is a senior research engineer of NTT Information Sharing Platform Laboratories. Previously, he was engaged in the development of end user computing environments. His research intersts includes distributed information systems and end user computing. Mr. Motoda received a B.S. and M.S. in information engineering from Toyohashi University of Technology in 1987 and 1989. He is a member of IPSJ.

Shuichiro Yamamoto is a senior research engineer, supervisor of NTT Information Sharing Platform Laboratories. Previously, he was engaged in the development of CASE tools and distributed application development platform. His research intersts includes distributed information systems, end user computing, and requirements engineering. Mr. Yamamoto received a B.S. in information engineering from Nagoya Institute of Technology in 1977, and M.S. in information engineering from Nagoya University in 1979. He is a member of IEEE, ACM, IPSJ, JSAI.