International Journal of Knowledge-based and Intelligent Engineering Systems 8 (2004) 149–160
IOS Press


# A system for generating user's chronological interest space from web browsing history

Harumi Murakami[a,*] and Takashi Hirata[b]
[a]*Graduate School for Creative Cities, Osaka City University, 3-3-138, Sugimoto, Sumiyoshi, Osaka 558-8585 Japan*
[b]*Japan Self Defense Force, Japan*

**Abstract**. We propose a method that helps users to understand their own interests by extracting terms from selected link texts and generating a new browsing history, and arranging those terms and Web-page icons on to the user's interest space in chronological order. We have implemented a prototype system based on this method. The system's performance was evaluated in two experiments, which revealed that (1) Japanese interest terms can be extracted from the user's selected link texts, (2) the interest-space browser displays the user's interest space on browsing the Web, and (3) the users can use this system to better understand their own interests, recall their past, and reorganize previously browsed Web pages.

## 1. Introduction

We often hear people say "Do what you like, and do what you are interested in." Many say that it is important to understand ourselves and do what we like for a sense of achievement or happiness. Psychologists have created numerous methodologies to help us understand ourselves. Vocational Preference Inventory (VPI) is one of such examples that help people to understand ourselves in terms of vocational preference. These methodologies are mainly based on a given set of questions. However, it is difficult to find something which is not written in the questions. In addition, methodologies using computers have not been thoroughly examined. The long-term goal of our research is to define methodologies of "understanding ourselves" using computers and develop systems using these methodologies.

As one element of this research, in this paper we aim to develop a system for understanding our interests by accumulating data related to our various activities on computers. That is, we examine Web browsing activities that take place in our daily lives.

*Corresponding author. Tel./Fax: +81 6 6605 3375; E-mail: harumi@media.osaka-cu.ac.jp.

Web browsers such as Microsoft Internet Explorer (hereafter IE) [1] or Netscape Navigator [2] have a "Web browsing history," which records URLs accessed by the user with dates and titles of the Web pages. Users can access previously accessed Web pages using this history.

It may be possible to develop a system to understand the user's interests by using this history. However, there are two problems. (1) Titles sometimes lack information on the contents of the Web pages. (2) Existing browsers only present a list of URLs of Web sites accessed one day and more ago, and it is difficult to determine what kinds of Web pages the user accessed before opening the URLs. In the former case, it may be more useful to see the "link text" or "anchor text" (hereafter, "link text") that the user selected than the title of the Web page. In the latter case, it may be more useful to see some keywords expressing the contents of the Web pages than the URL lists.

In this research, we focus on the user's link selecting activities and suggest two hypotheses. (1) It is possible to extract terms that express the user's interests from link texts selected by the users. (2) It is possible to develop a system to support users in understanding their interests by presenting these terms chronologically.

Based on the above hypotheses, we present a method of extracting terms from link texts and saving these

ISSN 1327-2314/04/$17.00 

terms as a new browsing history (hereafter "new history") with URLs and dates, arranging terms and "Web-page icons" that express Web pages in two dimensions, and presenting these terms and icons chronologically. We call the 2D space in which terms and Web-page icons are arranged "interest space," and the system that enables users to access previously browsed Web pages through the chronological interest space an "interest-space browser".

We have developed a new Web browser (hereafter "new browser") that generates a new browsing history and the interest-space browser, and performed two experiments.

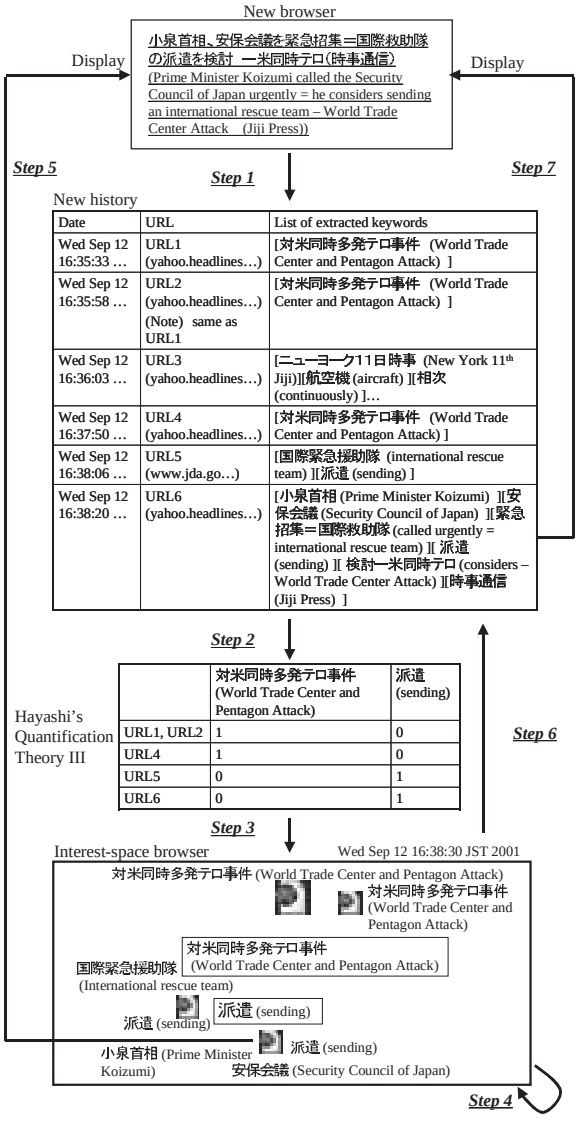
We will describe the proposed method in Section 2, and the experimental results in Section 3. Related work will be discussed in Section 4.

## 2. Method

### *2.1. Overview*

The proposed method is constructed from Step 1 to Step 7. Figure 1 shows an overview of the method.

- Step 1: When the user selects a link using a new browser, "date, URL designated by the link, terms extracted from a line including the link" are saved as a data set in a new history.
- Step 2: When the user executes "generate interest space" on an interest-space browser, the system treats "terms that occur more than twice in the new history (referred to as "interest terms") as category data, and treats URLs that have those interest terms as "sample data", and Hayashi's quantification theory III [3] (hereafter "Hayashi's theory III") is then applied to this data.
- Step 3: When the user executes "display interest space" on the interest-space browser, the interest space is displayed based on the results of Step 2.
- Step 4: When the user operates a "date bar" that changes the date and a "period bar" that changes the period of display, the display of the interest-space browser changes based on the date information.
- Step 5: When the user double-clicks Web-page icons that express Web pages on the interest-space browser, the new browser connects to the Internet and displays the designated Web pages.
- Step 6: When the user executes "open new history" on the interest-space browser, the system displays the list of the history for the current date.



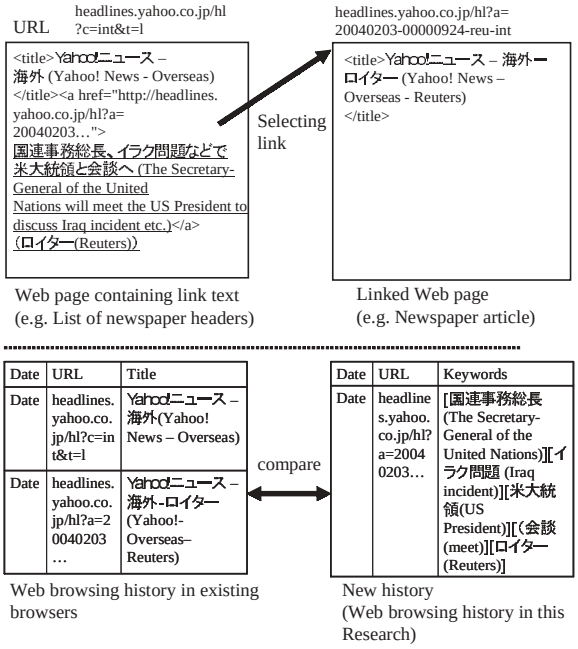

Note: The text in this example was originally Japanese and has been translated into English.

Fig. 1. Overview of the proposed method.

- Step 7: When the user double-clicks a piece of data in the history, the new browser connects to the Internet and displays the designated Web page.

### *2.2. Accumulating the new history on selecting links*

When the user selects a link on a new browser, date, URL designated by the link, and terms extracted from a line including the link are saved as data set in a new history.




Note: The text in this example was originally Japanese and has been translated into English.

Fig. 2. Comparison of histories in existing browsers and in the new browser used in this research.

Figure 2 shows the difference between the "history" of existing Web browsers and the "new history" in this research. The existing browsers save a date, a URL of Web page, and text extracted from title elements as a data set in the history. In contrast, our new browser saves a date, a URL designated by anchor elements, and terms extracted from a line including the selected link as a data set in the new history.

For example, consider that the user selects a link "国連事務総長、イラク問題などで米大統領と会談へ (The Secretary-General of the United Nations will meet with the US President to discuss Iraq incident etc.)" from "Yahoo! ニュース – 海外 – (Yahoo! Japan News – Overseas)". The history of the existing browser saves "Yahoo! ニュース – 海外 – ロイター – (Yahoo! News – Overseas - Reuters)," which is extracted from a title element of the Web page. Our new history saves "国連事務総長 (The Secretary-General of the United Nations), "イラク問題 (Iraq incident)," "米大統領 (US President)," "会談 (meet)," and "ロイター (Reuters)," which are extracted from a line including an anchor element of the previous Web page.

There are two methods for extracting terms from Japanese texts: (a) extracting mainly nouns using morphological analysis, and (b) extracting mainly non-hiragana characters by judging the characters. We use the latter method because it is easy to extract terms whose parts of speech are unknown.

In this research, we will focus on (1) extracting terms consisting of two or more characters (excluding hiragana or special characters), and (2) deleting unnecessary terms according to several heuristic rules and stop word lists created by the users.

The "judging unnecessary term" heuristic judges unnecessary terms from a text. We implement three kinds of heuristics: (a) numerical unnecessary term judgment, (b) date unnecessary term judgment, and (c) other unnecessary term judgment. For example, using (a), characters "10" "200" are judged as number unnecessary terms; using (b), characters consisting of "年 (year)" "月 (month)" "日 (day)" and numbers are date unnecessary terms; and using (c), characters ending with [的 (-like)] are judged as other unnecessary terms. The details of the heuristics are described in [4].

The user can select unnecessary terms displayed on the interest-space browser and record them in a stop word list.

In an example of Step 1 in Fig. 1, when the user selects a link "小泉首相、安保会議を緊急招集＝国際援助隊の派遣を検討ー米同時テロ (Prime Minister Koizumi called the Security Council of Japan urgently = he considers sending an international rescue team – World Trade Center Attack)," the following extracted terms are saved in the new history: "小泉首相 (Prime Minister Koizumi)," "安保会議 (Security Council of Japan)," "緊急招集 ＝ 国際援助隊 (called, urgently = international rescue team)," "派遣 (sending)," "検討ー米同時テロ (he considers – World Trade Center Attack)," "時事通信 (Jiji Press)" at URL6. If the user registers "時事通信 (Jiji Press)" in the stop word list, it will not be saved in the history.

### 2.3. Generating and displaying interest space

The user can select objects from "interest terms only" or "interest terms and Web-page icons" for display on to an interest space. This is intended to help users to explore the interest space quickly by selecting "interest terms only"; when the user finds the date of interest, he can select "interest terms and Web-page icons" to access the browsed Web pages.

When the user executes "generate interest space" after selecting objects to display in the interest-space browser, the system treats terms that occur two times or more in the new history (interest terms) as category

data and URLs that have interest terms as sample data, and calculates Hayashi's theory III.

Hayashi's theory III is a kind of principal component analysis for qualitative data that quantifiers binary elements, classifies samples or categories, and investigates characteristics of the data. By using Hayashi's theory III, similar data can be arranged close together: sample data with sample data, category data with category data, and sample data with category data.

An example of Step 2 in Fig. 1 shows that "対米同時多発テロ事件 (World Trade Center and Pentagon Attack)" and "派遣 (sending)" are treated as interest terms and category data, and that URL1, 4, 5, 6, which contain these terms, are treated as sample data. If the URLs are the same, they are merged as one URL. For instance, URL1 and URL2 are treated as URL1 and the frequency of URL1 becomes 2. Next, regarding URL1, "1" is entered in the "対米同時多発テロ事件 (World Trade Center and Pentagon Attack)" field because URL1 contains this term, and "0" is entered in the "派遣 (sending)" field because it does not contain the term.

When the user executes "display interest space", the interest space is displayed. The default date is the latest date in the new history and the default period is 30 days. Therefore, interest terms and URLs that occur within 30 days of the latest date are displayed. The 1st array becomes the X array and the 2nd array becomes the Y array, based on the results of calculation using Hayashi's theory III. The details of the calculation of 1st array and 2nd array are described in [4].

Interest terms are displayed as terms surrounded by squares. Web-page icons are displayed as icons of the earth. The Web-page icons change their size according to the frequency of URLs in the new history: large (three times or more), middle (twice), and small (once). They change their color according to the number of days counted from the current date: blue (three days or less), dark gray (seven days or less), and pale gray (more than seven days).

When the user sets a mouse cursor to a Web-page icon on the interest-space browser, extracted terms are displayed around the icon as "Web-page keywords" to help users to understand the contents of the Web page before opening the Web page.

An example of Step 3 shows that when the user executes "display interest space", "対米同時多発テロ事件 (Word Trade Center and Pentagon Attack)" and "派遣 (sending)" are displayed as terms in a square, and four different URLs are displayed as blue Web-page icons. The user can find Web pages referring to interest terms

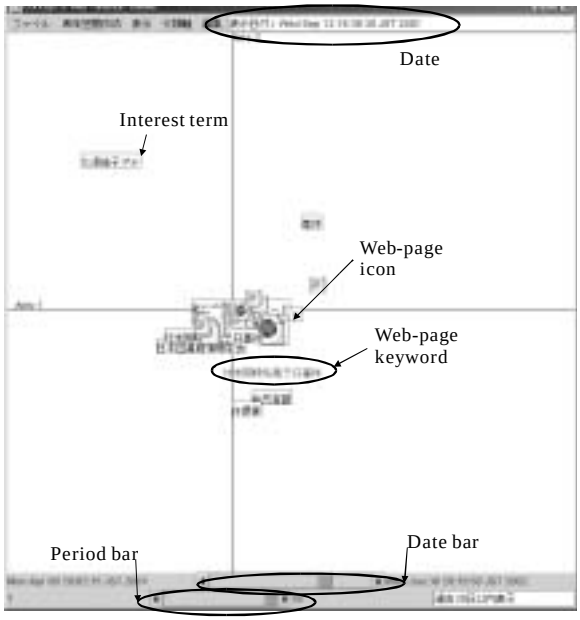



Fig. 3. An interest-space browser.

because the similar interest terms and Web-page icons are displayed close to one another. The user can find frequently browsed Web pages because the size of icons changes according to the frequency of selection. For example, the Web-page icons of URL1 are displayed larger than other Web-page icons.

Figure 3 shows a screen example of the interest space for the first author on 16:38:30, 12th of September, 2001, in Experiment 1. "対米同時多発テロ事件 (World Trade Center and Pentagon Attack)", "日本図書館情報学会 (Japan Association of Library and Information Science)", "久保純子アナ (TV announcer Ms. KUBO, Junko)", and "産休 (maternity leave)" are displayed as interest terms. The Web-page icon whose Web-page keywords are marked is URL1 and 2 of Fig. 1. The first author noticed that she often checked news about the World Trade Center and Pentagon Attack, recalled that she was writing for the annual meeting of Japan Association of Library and Information Science to be held in October, and noticed that she was interested in gossip about female TV personalities.

### *2.4. Displaying previously accessed Web pages*

There are two methods for displaying previously accessed Web pages from the interest-space browser.

One is as shown in Step 5 in Fig. 1. When the user selects a Web-page icon (URL6) referring to the Web-page keywords "小泉首相 (Prime Minister Koizumi)",

"安保会議 (Security Council of Japan)", and "派遣 (sending)" on the interest-space browser, a new browser will display the Web page of URL6.

Another is as shown in Steps 6 and 7. When the user executes "open new history", a list for the new history that centers on the current date will be displayed, so that the user can select URL6 and display the Web page.

## 3. Experiment

The implemented new browser has some display problems,[1] which are problems of implementation and not of algorithms.

We then judged that it is difficult to perform experiments in which users are asked to use the browser as though it were an existing browser, and planned two experiments described below.

- Experiment 1: The authors, who know the system very well, become subjects, perform simulations of using browsers in everyday life, and discuss problems and the utility of the method.
- Experiment 2: We limit Web sites in order that users can operate the implemented system, and discuss the usefulness of the method and the system.

### *3.1. Experiment 1*

#### *3.1.1. Method*

Subjects were the first author and the second author (hereafter Subject A and B). Subject A was a 37-year-old female, and a university lecturer in information science in Osaka, Japan. Subject B was a 31-year-old male, a member of the Japan Ground Self Defense Force, and a doctoral course student in information science.

They performed a simulation of using a new browser in everyday life, except for the following conditions.

- Subjects can use IE when an urgent need arises.
- When a Web page cannot be displayed, the page and the following selecting links can be displayed using IE.
- Subjects can use IE to display Web pages when they notice displaying problems.
- Subjects can use IE to display Web pages when they feel ashamed if the history is checked by the other subjects.
- Subjects generate and display the interest space every day that they use the system.

The period of the experiments were: April 9th to December 27th, 2001, for Subject A, and May 27th to November 13th, 2000, for Subject B. The experiments were done during working hours on weekdays in the laboratories, in principle.

#### *3.1.2. Results and discussion*

**Overview** For Subject A, the number of selected links was 1,485, with the following contents: Yahoo! Japan (yahoo.co.jp): 611 (41%); Osaka City University (osaka-cu.ac.jp): 94 (6%); Osaka University (osaka-u.ac.jp): 27 (2%); other: 753 (51%). The subject often viewed Yahoo! Japan and Web sites of universities in Osaka. The number of extracted terms was 2,488. The average number of terms per link was 1.68. The number of terms registered in her stop word list was 117.

For Subject B, the number of selected links was 808, with the following contents: Asahi Shinbun (www.asahi.com): 549(68%); Yahoo! Japan (yahoo.co.jp): 72 (9%); goo (goo.ne.jp): 28(3%); other: 159(20%). He viewed newspaper Web sites and the newspaper pages of Yahoo! Japan. The number of extracted terms was 2,112 (2.61 terms per link on average), and 748 terms[2] were registered in his stop word list.

**Extracting terms** We judged whether interest terms are adequate as words: about Subject A, 91% was adequate and 93% about Subject B.

Problems of extracting terms are classified into two groups, as follows.

1. Problems concerning recall rate in Japanese. (a) Terms including hiragana characters cannot be extracted (e.g. "つくば市 (Tsukuba City)"). (b) Terms consisting of one character cannot be extracted (e.g. "森 (forest)").
2. Problems concerning precision rate in Japanese. (c) Unnecessary symbols are attached (e.g. "結果 → (result→)"). (d) Symbols or parts of symbols are extracted (e.g. URL). (e) Terms tend to be long because of continuation of parts of speech. (f)

[1] For example, some web pages in particular using CGI and frames cannot be displayed, misaligned or garbled, or it takes a long time to display these pages.

[2] However, his stop word list was recorded for a different reason from this research and it is unclear how much of the stop words out of the 748 applied to this research.

Hiragana characters are cut from terms that end with hiragana characters (e.g. "見直(reexamine)".
(g) Terms are cut for unknown reasons.

Problems (a) and (b) cannot be solved by only the presented method by judging characters, and therefore morphological analysis or other methods should be added. Problems (c), (d), (e), and (g) can be solved by deleting unnecessary characters. Problem (f) is not major problem because users can understand the meaning of the terms.

**Extracting interest terms** Here, we consider how our method is useful in extracting interest terms by investigating ten high-frequency terms.

The top ten high-frequency terms for Subject A, without using her stop word list, were: "メッセージ (message)" (258), "最初 (first)" (130), "最新 (last)" (128), "一覧 (list)" (89), "..." (49), "ショッピング (shopping)" (42), "http" (39), "メッセージリスト(message list)" (39), "エンターテインメント (entertainment)" (32), "ホーム(home)" (32), and "旅行 (travel)" (32).

For Subject B, the top ten high-frequency terms without using his stop word list were: "Asahi" (229), "ニュース(news)" (147), "Yahoo" (76), "一覧(list)" (61), "テスト(test)" (59), "goo" (56), "ページ (page)" (33), "Mainichi" (32), "毎日テストデータ(Mainichi test data)" (24), and "スポーツ (sports)" (19).

These do not express the subjects' interests very well.

The top ten terms using her stop word list for Subject A contained the following terms: "日本 (Japan)" (22), "経済 (economy)" (12), "地域情報 (local information)" (12), "地方 (district)" (12), "コンピュータ(computer)" (11), "大学 (university)" (11), "旅行 (travel)" (10), "Java" (10), "教育 (education)" (10), "インターネット (Internet)" (9), "企業 (corporation)" (9), "研究 (research)" (9), "スポーツ (sports)" (9), and "ビジネス (business)" (9). These express the subject's interests as a university faculty member in the field of information science.

Top ten terms using his stop word list for Subject B were: "中国 (China)" (13), "北朝鮮 (North Korea)" (11), "森首相 (Prime Minister Mori)" (11), "ロシア (Russia)" (9), "アラファト議長 (Yasser Arafat)" (6), "イスラエル (Israel)" (6), "クリントン大統領 (President Clinton)" (6), "プーチン大統領 (President Putin)" (6), "首相(prime minister)" (5), "フィジー(Fiji)" (5), and "防衛庁長官 (Minister of State for Defense)"(5). These express Subject B's interests in international politics as a member of the military.

The above example suggests that high frequency terms include unnecessary terms, but these can be removed by using stop word list. Based on the results obtained for Subjects A and B, the percentage of unnecessary terms compared to extracted terms is projected to be between 5% and 30%.

Terms that do not express the user's interest are mainly terms included in link texts for Web pages that they often access, such as those included in the Yahoo! Japan message board (e.g. "メッセージ (message)", "最初 (first)", "最新 (last)" and those included in the Yahoo! Japan top page (e.g. "ショッピング (shopping)", "旅行 (travel)", and "音楽 (music)").

The above results suggest that the Japanese interest terms can be extracted from the user's selected link texts by investigating extracted high-frequency terms.

**Displaying interest space** Figure 3 shows that the interest space can be successfully displayed by using the presented method. However, the following problems were observed. (a) The interest space cannot be displayed when no term occurs two times or more. (b) Response is slow when the history is large. (c) Interest terms and Web-page icons including interest terms are often displayed close together, but either similar interest terms or similar Web-page icons are not displayed close together, intuitively. (d) It is difficult to identify the meaning of X and Y arrays. (e) Interest terms that occur one time, and Web-page icons that have no interest terms, cannot be displayed. (f) Interest terms and Web-page icons sometimes overlap and it is difficult to operate them.

Regarding (a) to (e), we need to improve the method of generating the history, or to try other displaying methods such as a self-organizing map. Concerning (b), creating a divided history file can be useful. For (f), scattering displayed objects should be introduced.

The results of Experiment 1 – for example, that the number of extracted terms for Subject B is larger than that of Subject A, and that the contents of the extracted terms for Subject B are more concrete than those for Subject A – suggest that better interest terms are extracted when users view newspaper Web sites. In addition, we found that the new browser can successfully display certain newspaper Web sites. Therefore, we planned an experiment targeting general users using newspaper Web sites (Experiment 2).

### *3.2. Experiment 2*

#### *3.2.1. Method*

**Subjects** The subjects were eight women, aged 24 to 57 years old (referred to as Subject A, B, C, ... H). Subject A, C, and F were housewives, Subject B was

a systems engineer, Subject D was a secretary, Subject E and H were undergraduates students, and Subject G was an office worker.

**Browsing** We downloaded newspaper articles (HTML files) over a three-month period (June to August, 2000) from the newspaper Web site "Mainichi Interactive [5]" The subjects were asked to browse any article they wanted to see, at least one page per day.

In the Mainichi Interactive site, "a line including a selected link" is the same as the link text, and is also the same as the article header. The title of the article is one such as "毎日インタラクティブ・記事全文 (Mainichi Interactive Full-text Article)." It contains no information about contents of the article. In this case, it is evident that link texts are better than titles as resources for term extraction.

Here, we investigate which resource is better: link texts (i.e. headers of articles) or Web pages (i.e. contents of articles including headers).

After browsing the article, we asked the subjects to evaluate interest terms as follows.

1. To investigate which more effectively expresses interests of users–term extracted from link texts or those extracted from Web pages–we showed two lists of "extracted terms ordered by frequency" from both link texts and Web pages, and asked the subjects the following question: "These two lists are interest terms using a different method from your browsing activities. Which list do you think intuitively to express your interests?"
2. To investigate whether "extracted terms from link texts that occur two times or more" are appropriate for interest terms, we asked the subjects to evaluate each extracted term that occurs two times or more in the light of "Do you think the term expresses your interests in your browsing?" The evaluation is categorized as follows: 5: agree strongly, 4: agree moderately: 3: neutral; 2: disagree moderately; and 1: disagree strongly. The average of evaluation values of each term is regarded as "interest rate X" in link texts. In contrast, we also asked subjects to evaluate terms extracted from Web pages whose number was the same as the number of terms with a frequency of two or more extracted from link texts. The average of these evaluations is considered interest rate X for that Web page.[3]
3. To investigate the general quality of the high-frequency interest terms extracted from link texts and Web pages, we asked subjects to conduct an evaluation as described in 2. above for the ten terms with the highest frequency, and took this value as interest rate Y.
4. It seems that the quality level of interest terms is different. To investigate which is better for resources to extract interest terms of good quality, we asked the subjects to select the ten terms that best express interest terms with the greatest frequency from extracted terms for interest rate X.

**Evaluation of interest space browser** To investigate the usefulness of the interest-space browser, after we asked subjects to "generate interest space" and to "display interest space", we gave them the following instructions.

> The displayed window automatically shows a space generated from your Web browsing history. Terms surrounded by a square are terms extracted as your interest terms.
> An "earth" icon shows a Web page you browsed. Gray icons shows past browsing, and larger icons indicate the frequency of the browsing.
> Blue text displayed on an icon when you move the cursor shows terms extracted from the Web page.
> There are two slide bars at the bottom of the window, a date bar and a period bar. Please move the date bar from left to right; this will change the display from past items to current items. Move the period bar as you like, and watch the window.
> Please move objects (terms and icons) using the mouse when they are difficult to see.

Four questions were asked: (Q1) Does the interest space indicate your interest space when you actually browsed Web pages? (Q2) Is the system useful for understanding your interests? (Q3) Is it useful for reorganizing Web pages?, and (Q4) Is it useful for recalling your past?

The evaluations were categorized as follows: 5: agree strongly; 4: agree moderately; 3: neutral; 2: disagree moderately; and 1: disagree strongly. The subjects were asked to describe reasons for each evaluation.

The experiment was done in December 2000 for Subjects A-D, and from July to December 2001 for Subjects E-H.

[3] It was possible to ask subjects to evaluate terms that occur two times or more extracted from web pages, but the number of terms was enormous and the difference in the burden on subjects was large, so we did not adopt this method.

Table 1
Rates of interest for extracted terms

| | Link texts | Web pages |
|---|---|---|
| Terms | 1,524 (SD = 842) | 11,209 (SD = 6,870) |
| Selected links | 335 (SD = 256) | As left |
| Terms/Selected links | 3.63 (SD = 0.19) | 33.49 (SD = 3.77) |
| Interest rate X | 3.30 (SD = 0.52) | 3.22 (SD = 0.47) |
| Interest rate Y | 3.89 (SD = 0.34) | 3.19 (SD = 0.47) |

Note: Interest rate X for link texts: average of evaluation of extracted terms that occur two times or more; Interest rate X for Web pages: average of evaluation of extracted terms whose number is the same as the number of terms with a frequency of two or more extracted from link texts; Interest rate Y: average evaluations for ten interest terms that appeared with the highest frequency.

### *3.2.2. Results and discussion*

**Overview** Table 1 lists the results: 335 links were selected, 3.63 interest terms per link were extracted from link texts, and 33.49 were extracted from Web pages on average.

**Evaluation of extracting interest terms** The subjects were asked whether terms from link texts or from Web pages matched their interests more closely. All subjects said that link texts were more helpful.

Table 2 shows the most frequent extracted term from link texts and Web pages. It can be seen at a glance that in the case of high-frequency terms extracted from link texts, there were more terms expressing proper names and current topics than in the case of Web pages, and that it is easier to tell where the user's interests lie.

Interest terms occurring two times or more were evaluated for each subject. For link texts, the average score was 3.30, while for Web pages the average score was 3.22 (Interest rate X)[$t = 0.19$, $df = 7$, ns]. Average evaluations for ten interest terms that appeared with the highest frequency were 3.89 for link texts and 3.19 for Web pages (Interest rate Y). An analysis of variance revealed a significant effect for link texts [$t = 2.82$, $df = 7$, $p =< 0.05$].

Interest rate X is 3.30 and Interest rate Y is 3.89 in link texts. Another analysis of variance showed a significant effect for Interest rate Y in link texts [$t =$ 3.05, df $= 7$, $p =< 0.05$]. Interest rate X is 3.22 and Interest rate Y is 3.19 in Web pages [$t = 0.19$, df $= 7$, ns].

We asked the subjects to select the ten terms in which they had the most interest from among the group of terms with interest rate X. The terms that held the most interest were selected from among the link texts by all subjects, and all terms were in the top one or two in terms of frequency rate.

From the above results, we confirmed that in the case of newspaper articles, it is better to target link texts than Web pages for interest terms, and to target high-frequency terms rather than low-frequency terms for interest terms when using a simple algorithm based on term frequency. We believe that this indicates the merits of a method that focuses on links, as described in this research, and the effectiveness of the proposed method. Furthermore, we found that when only three months worth of newspaper articles are used as the information source, interest terms can be extracted from high-frequency terms without the use of stop word lists.

**Evaluation of interest-space browser** Table 3 shows the results of the evaluation. We consider the results positive.

Concerning display of interest space (Question 1), the results were positive. Subject B wrote "Terms express my interest well." Subject D wrote "I can easily find what I saw many times." These comments were about interest terms. Negative comments include "I couldn't see the meaning of the space" (Subject B).

Among the comments about understanding of self interests (Question 2), "I recognized my interests and I could recall what I was interested in" (Subject G) and "I notice biased thought" (Subject G) were positive, but the comment "I cannot see the classification and I don't know how the system can be useful" (Subject E) was negative.

About reorganization of Web pages (Question 3), positive comments suggest the possibility of reorganizing Web pages according to users interest, like "If I am very interested in the contents and I want to recall them later, the system should be useful" (Subject B) and "I think it is easy to reorganize Web pages by using the system because I can find what I am interested in and I can find Web pages that I saw many times" (Subject C). However, some problems were pointed out, such as "I cannot find the meaning of the space" (Subject A) and "It is difficult to view the display" (Subject H).

About recalling the user's past (Question 4), comments such as "The system is useful for recalling past interests and news" (Subject H) and "I can recall what I was interested in before, but I'm not sure that I can use this to remember my past" (Subject H) suggest that the system is useful in recalling past interests and newspaper articles.

In Experiment 2, it was suggested that the interest-space browser displays the user's interest space on browsing the Web, and that the user can understand their interest, recall their past, and reorganize previously browsed Web pages by using the system. We believe that these comments indicate that it is possible to develop a system for supporting an understanding of user interests by targeting newspaper article Web sites.

Table 2
The highest frequency extracted interest terms

| Subjects | Link texts | Web pages |
|---|---|---|
| A | "サミット (Summit)" (3), "シドニー五輪 (Sydney Olympic) (3)", "NY 株 (NY stock)" (3), "パソコン (PC)" (3) | "見込 (prospect)" (11) |
| B | "[シドニー五輪 (Sydney Olympic)]" (7) | "確認 (confirmation)" (28) |
| C | "[露原潜事故 (Russian nuclear submarine accident)]" (22) | "死亡 (death)" (112) |
| D | "逮捕 (arrest)" (14) | "死亡 (death)" (47) |
| E | "偽造 (forgery)" (7), "[爆発 (explosion)]" (7) | "製造 (manufacture)" (30), "調査 (investigation)" (30) |
| F | "[異物混入 (tampered drink)]" (17), "自主回収 (voluntary recall)" (17) | "製造 (manufacture)" (74) |
| G | "[訃報 (news of someone's death)]" (6) | "自宅 (at home)" (20) |
| H | "[サッカー (soccer)]" (25) | "説明 (explanation)" (58) |

Note: Terms in brackets [ ] are terms of greatest interest to Subjects.

Table 3
Evaluation for an interest-space browser

| Questionnaire | Mean | SD |
|---|---|---|
| (Q1) Does the interest space display your interest space when you actually browsed Web pages? | 4.13 | 0.83 |
| (Q2) Is the system useful for understanding your interests? | 4.25 | 0.50 |
| (Q3) Is it useful for reorganizing Web pages? | 3.75 | 0.71 |
| (Q4) Is it useful for recalling your past? | 4.00 | 0.93 |

## 4. Related work and discussion

The prototype system described in this paper is one of components of a system called Memory-Organizer [6], which accumulates human externalized memory on computers. Memory-Organizer provides users with an integrated environment for intelligent information activities; not only Web browsing, but also integrating bookmarks and memoranda, overwriting Web pages, and reusing search results of search engines. This paper described a method for generating a new history and an interest-space browser in Memory-Organizer.

Much research concerning visualization of Web browsing history deals with existing history such as IE, and there is not much research on methods for generating new history. The typical approach is that "visualization of history helps users in navigating the Web," in which Web pages or sites are expressed as "nodes" and the relation between nodes is displayed as a tree or graph structure. Examples of these can be seen in MosaicG [7] and PadPrints [8]. This approach is useful for "going back" to recently browsed Web pages, as in the case of existing browsers. However, it is unclear whether this approach is useful for displaying users' interests, and for helping people to access browsed pages they viewed more than one month ago.

WebWatcher [9] and Letizia [10] learn user's interests for navigating and information gathering; in other words, the goal of these systems is to help users access Web pages they might be interested in in the future. However, they do not present what the users saw visually.

Although different to visualization of browsing histories, there are systems that categorize similar Web pages or sites and display these groups on a map along with terms (e.g., WEBSOM [11], which uses self-organization). The map display can also express the user's interests. The question of which method is more effective–a map display or a 2D space display–is one that will require further investigation in the future.

There has been research into providing support for information searches, communication, and the formation of ideas by expressing the user interest space in a 2D space (e.g., [12]). This is similar to the current research in that user interest is expressed in a 2D space. However, our work differs in terms of the goals of the research (i.e. understanding the user's interests in our work), the information source (i.e. Web browsing history), and the method of display (i.e. displaying the interest space chronologically).

According to the well-known "Johari's Window," there are four regions in human mind; A: a region in which things are known by the self and others, B: a region in which things are known by others but not by the self, C: a region in which things are known by the self but not by others, and D: a region in which things are

unknown by the self and others. It is considered that in a well-developed personality, A should be larger, and D should be smaller.

In which region can this system help users to understand their interests? Some comments such as "Terms express my interest well" (Subject B) or "I can recognize what I am interested in" (Subject H) in Experiment 2 suggest that the system can help mainly in the region that the user already knows. However, as the subjects knew that the history would be checked by experimenters, the region in which the subjects intended to hide their interests is unclear. The comment "I notice biased thinking" (Subject G) suggests that the system can help in regions that the user does not know.

In this research, we have set two hypotheses. (1) It is possible to extract terms that express user's interests from the link texts. (2) It is possible to develop a system to support users in understanding their interests by presenting the terms chronologically. We think that hypothesis (2) has been proved by two experiments. Hypothesis (1) has been partially proved in the case of Japanese, with some limitations. We will describe future work below.

## 5. Future work

As discussed above, due to problems involved related to Web browser implementation, we were unable to conduct experiments using the system in a natural setting targeting all types of Web sites. The practical problem revolves around the extremely time-consuming work involved in installing the system in a Web browser on a practical application level, an operation that falls outside of the scope of this research. Below, we will discuss the issues that can be verified with the framework of this research.

### *5.1. Resources for extracting interest terms*

We have found that it is better to extract interest terms from link texts than Web pages by using a simple method based on the term frequency in a newspaper Web site. However, there are other methods for extracting terms from Web pages, and we need to compare the results. This experiment did not cover all types of Web pages, and it is thus unclear as to whether extraction of terms from link texts is always better than extraction from Web pages.

The intent of this research is not to attempt to replace existing Web browsing histories, but rather to add "information extracted from link selection behavior" to existing Web browsing histories. In the future, rather than conducting extensive investigations into which text approach is better, we intend to focus our attention on studying methods for handling both types of text effectively.

### *5.2. Term extraction method*

We adopted a character-based method for extracting terms from text. In the future, in order to resolve the problems that became apparent in Experiment 1, we will study ways of improving upon the current method, and of combining this method with morphological analysis.

### *5.3. Coverage of interest terms*

In the interest-space browser, we were unable to examine terms with a frequency of one, or Web pages with no interest terms. It would be possible to compensate for this using a function for opening the history itself making reference to the interest-space browser, but we feel that an even better approach would be to resolve this problem within the framework of the interest-space browser method. We will pursue this as a research theme in the future.

### *5.4. Display method and user-interface*

There are problems involved in using an interest space display method based on a combination of the new history and Hayashi's theory III, and it will be necessary to investigate improvements to the history creation method and trials for space arrangement methods other than Hayashi's theory III. Regarding the problem of objects being grouped together in the center, we must examine improvement measures including appropriate scattering methods; regarding the problem of the time required to display results when the volume of history increases, we must study methods such as calculating histories in segments.

## 6. Conclusions

We proposed a method that helps users to understand their own interests by extracting terms from selected link texts and generating a new browsing history, and then arranging those terms and Web-page icons in the user's interest space in chronological order.

The features of this method are: (a) the creation of a new history based on the user's link selecting activities, and (b) a displaying method that arranges terms extracted from the user's link selecting activities, along with Web-page icons, in a two-dimension space, and presents these results chronologically.

We have implemented a prototype that consists of a new browser and an interest-space browser.

The system's performance was evaluated in two experiments, which revealed that (1) Japanese interest terms can be extracted from the user's selected link texts, (2) the interest-space browser displays the user's interest space on browsing the Web, and that (3) the user can use this system to better understand their own interests, recall their past, and reorganize previously browsed Web pages.

There are areas requiring improvement in each of the algorithms, including those for interest term extraction methods and display methods, which will be adopted as research themes for the future. We will also conduct studies focusing on different information sources.

## Appendix: Hayashi's Quantification Theory III

Hayashi's Quantification Theory III is a method for quantifying elements appearing in the form of binary data or frequencies, sorting samples and categories, and using this data to investigate characteristics.

Let us assume that M is a series of binary elements.

$$M = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1m} \\ d_{21} & d_{22} & \dots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{nm} \end{bmatrix}$$

The sums of each column $c_j$ and the sum of each row $r_j$ are derived. At this time, if $n < m$, the rows and columns are reversed.

$$c_j = \sum_{i=1}^{n} d_{ij} (j = 1, 2, \dots, m),$$

$$r_j = \sum_{j=1}^{m} d_{ij} (i = 1, 2, \dots, n)$$

Next, a diagonal matrix is created using $(c_j)^{-1/2}$ and $(r_j)^{-1}$ as the respective constitutional elements.

$$C = \begin{bmatrix} \frac{1}{\sqrt{c_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{c_2}} & \dots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{\sqrt{c_m}} \end{bmatrix} R = \begin{bmatrix} \frac{1}{r_1} & 0 & \dots & 0 \\ 0 & \frac{1}{r_2} & \dots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 0 & \dots & 0 & \frac{1}{r_n} \end{bmatrix}$$

Matrix $D$ is calculated as follows, from matrixes $C$ and $R$ and the original matrix $M$.

$D = CM^T RMC$($M^T$ is a transposed matrix of $M$)

Eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_{m-1}$ for Matrix D are derived. At this time, elements with an eigenvalue of 1 are omitted.

$$D - \lambda^2 E = 0 \ (E \text{ is an unit matrix})$$

Furthermore, let us assume that $V$ is the matrix whose columns are comprised of eigenvectors $v_1, v_2, \dots, v_{m-1}$ corresponding to each eigenvalue.

Finally, category quantity $X$ is calculated as:

$$X = CV$$

and sample quantity $Y$ is calculated as:

$$Y = \frac{1}{\lambda} RMX$$

**Category data** A value that corresponds to the factor loading for principal component analysis. It is possible to infer by analogy the relationship of category variables using only a magnitude relation between these values, but by selecting two desired category quantities and drawing a distribution graph, it is also possible to study the relationship between those two categories. Using this plot, it is possible to infer by analogy what characteristics the category quantities themselves extracted from the category variables.

**Sample data** A value that corresponds to the principal component value in principal component analysis. It is possible to infer by analogy the relationship of category variables using only a magnitude relation between these values, but by selecting two desired sample quantities and drawing a distribution graph, it is also possible to grasp the relationship between those two samples. By combining this plot with the category quantity plot, it is possible to learn how the characteristics of the category variables are reflected in the sample.

## References

[1] http://www.microsoft.com/.

[2] http://www.netscape.co.jp/.
[3] E. Kinoshita, *Wakariyasui-Suugaku-Model-Ni-Yoru-Tahen ryokaiseki-Nyumon* (in Japanese), (Introduction to Comprehensible Multi-variant analysis), Keigakushuppan, 1987.
[4] T. Hirata, *Studies on the Support for Constructing and Sharing Externalized Memory* (in Japanese), D. Eng. Thesis, Graduate School of Information Science, Nara Institute of Science and Technology, 2001.
[5] http://www.mainichi.co.jp/.
[6] H. Murakami and T. Hirata, Information Acquisition and Reorganization from the WWW by using Memory-Organizer, *Bulletin of Osaka City University Media Center* **3** (2002), 9–24.
[7] E.Z. Ayers and J.T. Stasko, *Using Graphic History in Browsing the World Wide Web*, Proceedings of WWW4, 1996.
[8] R.R. Hightower, L.T. Ring, J.L. Helfman, B.B. Bederson and J.D. Hollan, *Graphical Multiscale Web Histories: A Study of Padprints*, in Proceedings of ACM Hypertext'98, 1998, pp. 58–65.
[9] R. Armstrong, D. Freitag, T. Joachims and T. Mitchell, *WebWatcher: A Learning Apprentice for the World Wide Web*, Proceedings of AAAI Symposium on Information Gathering from Distributed, Heterogeneous Environments, 1995, pp. 6–12.
[10] H. Lieberman, *Letizia: An Agent That Assists Web Browsing*, Proceedings of IJCAI95, 1995, pp. 924–929.
[11] T. Kohonen, S. Kaski, J. Lagus, Sakojvi, V. Paatero A. and Saarela, Self Organization of a Massive Document Collection, IEEE Transactions on Neural Networks, *Special Issues on Neural Networks for Data Mining and Knowledge Discovery* **11**(3) (2000), 574–585.
[12] R. Kadobayashi, K. Nishimoto, Y. Sumi and K. Mase, Personalizing Semantic Structure of Museum Exhibitions by Mediating between Curators and Visitors (in Japanese), *IPSJ Journal* **40**(3) (1999), 980–989.