Information Acquisition and Reorganization from the WWW by using Memory-Organizer

MURAKAMI Harumi[†] and HIRATA Takashi[‡]

We have developed a system called Memory-Organizer that helps users to construct "externalized memory" and supports their creative activities. The system consists of (a) a thinking-space browser, which helps users to input and edit externalized memory and support creative thinking; (b) an overlay web browser, which helps users to gather, clip, and create externalized memory while browsing the Web; and (c) an interest-space browser, which helps users to recall externalized memory by arranging it on the user's interest space in chronological order. We used Memory-Organizer for information acquisition and reorganization while browsing the Web. It's performance was evaluated in two experiments, which revealed that (1) the algorithm of extracting interest terms from web pages works well and that (2) the interest-space browser can successfully display the user's interest space, and the user can recall their past by using this browser and reorganize previously browsed web pages of newspaper articles accordingly.

Keywords: Memory-Organizer, externalized memory, overlay web browser, thinking-space browser, interest-space browser

Memory-Organizer を用いた Web からの情報獲得・整理

村上 晴美[†] 平田 高志[‡]

個人の「外化記憶」を構築し, 知的生産活動を支援するシステム Memory-Organizer を試作した. 本システムは, (a) ユーザのアイディアから外化記憶の作成・編集を支援する思考空間ブラウザ, (b) ユーザの Web 閲覧時の外化記憶の作成, 収集, 抽出を支援するオーバレイ Web ブラウザ, (c) ユーザの時系列の興味空間上に外化記憶を配置することにより外化記憶の想起を支援する興味空間ブラウザから構成される. 本稿では WWW からの情報獲得, 整理に焦点をあてる. ユーザの Web ブラウジング履歴から個人の興味空間を生成して Web ページを整理する実験を行ったところ, 興味空間ブラウザがユーザの興味空間を表していることや, 過去の想起や Web ページの整理に役立つ可能性があることがわかった.

キーワード: Memory-Organizer, 外化記憶, 思考空間ブラウザ, オーバレイ Web ブラウザ, 興味空間ブラウザ

1 Introduction

The Internet has permeated our daily life. Much research has been done on information gathering, retrieval, and organization from the Internet. However, most of this research focuses on new techniques and algorithms; there has not been much research on practical systems to meet simple needs of individuals. For example, a user may have the following requirements: (a) clipping information from web pages, (b) overwriting memoranda on web pages, (c) integrating one's ideas and information on the Web, and (d) recalling browsed web pages from bookmarks or history files more easily.

To meet the above requirements, we have devised the idea of constructing "externalized memory." We use the term as an externalized human memory built onto a computer. By externalizing human memory and integrating it with various information sources, we aim to develop an information environment that facilitates human creative activities.

We have developed a system called Memory-Organizer that helps users to construct an externalized memory. Memory-Organizer consists of (a) a thinking-space browser, which helps users to input and edit externalized memory and support creative thinking; (b) an overlay web browser, which helps users to gather, clip, and create externalized memory while browsing the Web, and

[†] Media Center, Osaka City University
大阪市立大学 学術情報総合センター
[‡] Japan Ground Self Defense Force
防衛庁 陸上自衛隊

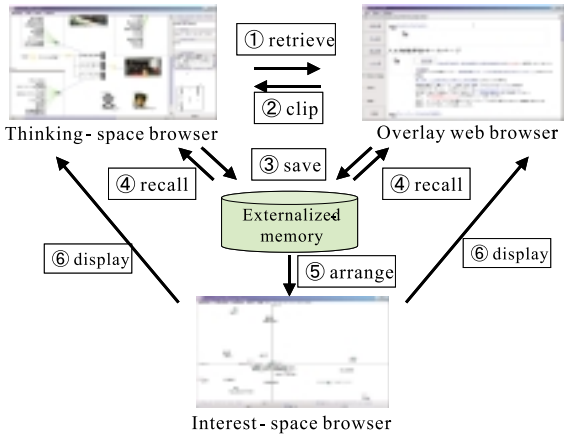

Figure 1: Overview of Memory-Organizer.

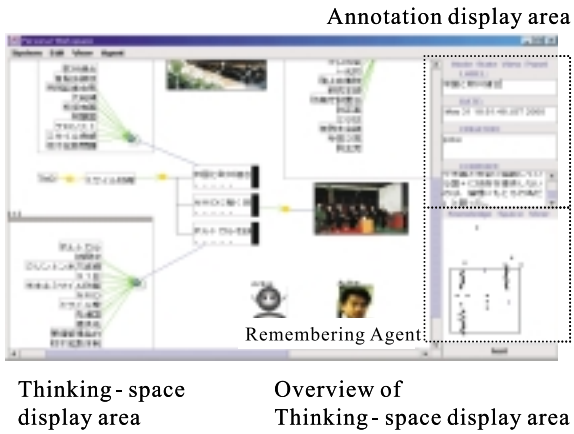

Figure 2: A thinking-space browser.

(c) an interest-space browser, which helps users to recall externalized memory by arranging it on the users' interest space in chronological order. An overview of the system is shown in Figure 1.

2 Thinking-space browser

The thinking-space browser helps users to input and edit externalized memory and support creative thinking.

2.1 Creating and recalling units

We call basic units of externalized memory "units." There are three kinds of units: (a) a concept unit, which expresses concepts like keywords; (b) an external-data unit, which points to external files such as texts and images; and (c) a relation unit, which displays relations between units. Currently, we deal with texts, images, and web pages as external-data units. They are called "external-text unit", "image unit", and "web-page unit", respectively. An associative relation (hereafter referred to an "association unit") is major among relation units. The starting point of the association unit is called a "key unit" and the terminal points is called a "value unit." The association unit is a single link that connects one or many key units with one or many value units that represents memories triggered by the given key units. The notation used for the association unit is borrowed from that of associative representation in CoMeMo [1, 2], which is a previous version of the Memory-Organizer.

Figure 2 shows a screen image of the thinking-space browser. The thinking-space browser consists of three areas: a thinking-space display area, an annotation display area, and an overview of the thinking-space display area. The user can input, edit, and recall units by mouse or pen. The system enables users to recall externalized memory by using units displayed on the thinking-space browser, the global position of the users, and date information concerning the externalized memory. A remembering agent helps users to recall externalized memory by monitoring the display of the system, which displays similar units when a new unit is created.

2.2 Web search

The user can send a request to search engines via an overlay web browser by using terms displayed on the thinking-space browser. When more than two concepts are selected, the system combines them with "And" and does a "simple search" by Yahoo! Japan (www.yahoo.co.jp). Then the search results are displayed, and the user can find information by browsing the Web on the overlay web browser. The user can clip information on the overlay web browser onto the thinking-space browser as described in section 3.2. The user can send another request to search engines by using terms displayed on the thinking space browser. This seamless "search and clip" processes helps the creative activities of users.

3 Overlay web browser

The overlay web browser helps users to gather, clip, and create externalized memory while browsing the Web. Two main functions of the browser are (1) to overwrite information on a displayed web page, and (2) to clip information from it.

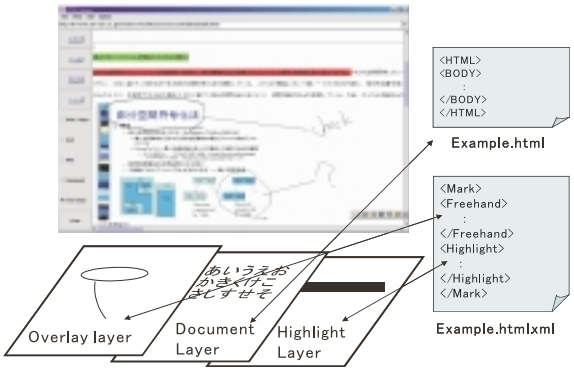

Figure 3: An overlay web browser.

3.1 Overwriting information on web pages

The user can overwrite information (e.g., highlights and memoranda) on displayed web pages.

Figure 3 shows an overview of the overlay web browser. The browser is constructed of three layers: a document layer, which shows an original web page; an overlay layer, on which the user writes a memorandum by freehand; and a highlight layer, to which the user attaches highlights. The document layer is saved in a HTML file, and the overlay and highlight layers are saved together in an XML file. The XML file is linked with the original HTML file by name.

3.2 Clipping information from web pages

The user can clip information (e.g., texts and images) on an overlay web browser onto the thinking-space browser. When the user traces a piece of text (characters), (1) keywords are extracted from the text according to the algorithm given below and concept units are created, (2) an external-text unit is created from the text itself, and (3) a web-page unit - which links these concept units (as key units) with the external-text unit (as value units) - is created.

Figure 4 shows the simple algorithm used for clipping information. To reduce the time of extracting keywords, we use quite a simple algorithm that has two functions: (1) extracting terms longer than two characters (except for hiragana or special characters) and (2) deleting unnecessary keywords according some several heuristics rules and unnecessary term lists created by the users. Some examples of the heuristics are "deleting a keyword which has the term 「的」(-like) at the end" and "discarding characters that can be regarded as dates."

Web-page units are displayed with an icon called

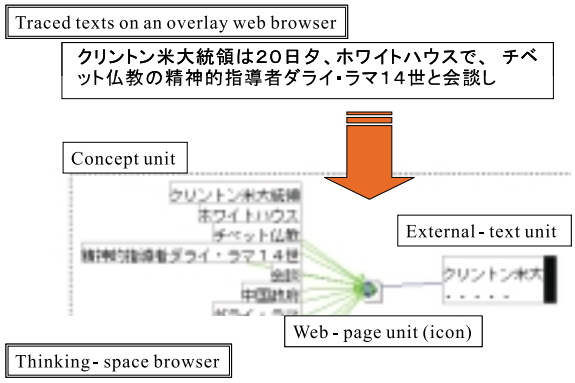

Figure 4: Clipping information from web pages.

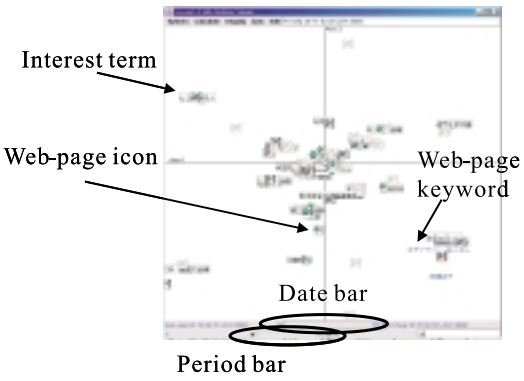

Figure 5: An interest-space browser.

a "web-page icon." If the user double-clicks a web-page icon, the overlay web browser connects with the Internet and displays the web page.

When the user single-clicks the image on the overlay web browser, the image file is downloaded and an image unit is created.

4 Interest-space browser

The interest space browser helps users to recall externalized memory by arranging it on the users' interest space in chronological order.

The interest space is generated according to the similarity of keywords extracted as interest terms. The user can see what they are interested in and access externalized memory.

Figure 5 shows a screen image of the interest-space browser. An interest term is displayed on the interest space browser as a keyword surrounded by a square, and a web page is displayed by a web-page icon. When the user double-clicks the web-page icon, the overlay web browser displays the selected web page. When the user moves the mouse cursor onto a web-page icon, interest terms concerning the corresponding web page are displayed, thereby allowing the user to get an idea

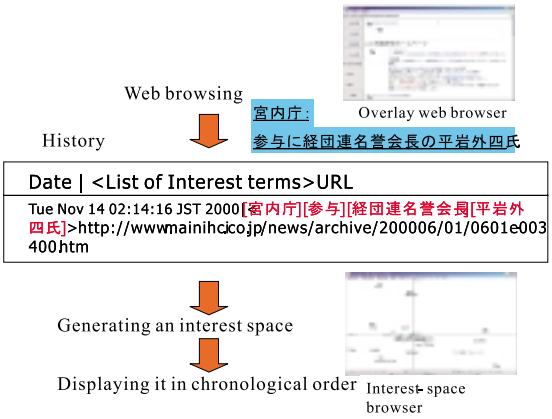

Figure 6: Generating an interest space.

of the contents of the web page before opening it. The user can take a look at the interest space displayed in chronological order. The user can select a "date" from the times they used the browser by using the date bar, and select a "period" from 1 to 30 days by using the period bar. A web-page icon changes its color according to the date the page was browsed: blue (less than three days), dark gray (less than one week), and pale gray (over a week). It also changes its size according to the number of times the page was browsed: large (three times or more), medium (two times), and small (one time). These functions help users to recall previously accessed web pages more easily.

The following describes how interest terms are extracted when the user browses the Web by using the overlay web browser. When the user single-clicks a link on the overlay web browser, keywords are extracted from the selected link texts automatically. Currently, the system does not extract any keyword from the document. To improve precision rate of interest terms and to reduce the time of extraction, we use only link texts, not documents. The algorithm for the extraction is as the same as that described in section 3.2. Extracted keywords are saved in a history file (index file) with the date the user browsed the web page and its URL. Currently, Hayashi's quantification theory III[3] is used to calculate the similarity between keywords and URLs and display the extracted keywords and web-page icons. Figure 6 shows the process of generating the interest space.

5 Experiment

We evaluated the usefulness of the overlay web browser and the interest-space browser in two separate experiments.

5.1 Experiment 1

Method We investigate how the second author (male, 30 years old, the system creator) used the overlay web browser and the interest-space browser. The experiment was performed from 26 May 2000 to 31 Jan 2001 (around eight months).

Results and discussion 2,545 interest terms were extracted, and 1,052 web pages were browsed. The web sites browsed were asahi.com (www.asahi.com[1]), 757 times (72.0%); Yahoo! Japan (www.yahoo.co.jp[2]), 74 times (7.0%); goo (www.goo.ne.jp[3]), 30 times (2.9%); and others, 191 times (18.2%). The system extracted 2.42 interest terms per web page. There were 715 unnecessary words found registered in the user's unnecessary list. We checked 2,545 interest terms according to whether they were appropriate as words. Consequently, we found (1) 2,403 as appropriate (94.4%), (2) 119 appropriate if edited (4.7%), (3) 23 as inappropriate (0.9%). Although the extraction method is quite simple, we analyze it to be useful. The main fault of this method is that it is unable to extract keywords that contain a Japanese "hiragana" character. The most frequent term was "中国 (China)", and the second was "北朝鮮 (North Korea)", and the third was "森首相 (Prime Minister Mori)." We found that the high-frequency terms reflect the user's interests, such as international relationships and politics, that are related to his work.

Experiment 1 showed that the system was most suitable for browsing newspaper web site articles so we chose to focus on newspaper articles in Experiment 2.

5.2 Experiment 2

5.2.1 Method

Subjects were four womens, aged 23 to 37 years old (referred to as A, B, C, and C). We downloaded newspaper articles (HTML files) over a three-month period (June to August, 2000) from a newspaper web site (www.mainichi.co.jp). The subjects were asked to browse any article they wanted to see, at least one page per day. After browsing the article, we asked the subjects to evaluate interest terms and the system. The evaluations were categorized as follows: 5: agree a lot; 4: agree moderately; 3: neutral; 2: disagree moderately; and 1: disagree a lot. Two tests were conducted: displaying only interest terms (Test 1) and displaying interest terms and web-page icons (Test 2). Five questions

[1] One of the most popular newspaper sites in Japan
[2] The most popular search site in Japan.
[3] One of the most popular search sites in Japan.

Table 1: Evaluation of the interest term.

	A	B	C	D	Mean
Web pages	109	240	902	236	372
Keys/P(L)	3.82	3.61	3.37	3.69	3.62
Keys/P(D)	39.43	36.62	28.08	31.31	33.86
Interest(L)	3.85	4.08	3.73	3.50	3.79
Interest(D)	3.50	3.33	3.36	3.00	3.30

Web pages: number of browsed web pages.
Keys: number of extracted keywords.
L(Link): extracted from link texts.
D(Document): extracted from documents.
Interest: interest term rate.
Interest(L): average of evaluation of interest terms occurring more than twice
Interest(D): average of evaluation of interest terms whose number is equivalent of interest(L).

were asked in each test: Q1: Does the interest space display your interest space when you actually browsed web pages?; Q2: Do you want to see the interest space in the future?; Q3: Is the system useful for helping you learn about yourself?; Q4: Is it useful for reorganizing web pages?; and Q5: Is it useful for recalling your past?. The experiment was done in December 2000.

5.2.2 Results and discussion

Two types of texts as an information source for extracting interest terms were evaluated: (a) link texts, or headers of a newspaper article (as in Experiment 1), and (b) documents, or articles. Table 1 lists the results: 372 web pages were browsed, 3.62 interest terms per web page were extracted from link texts, and 33.86 were from documents on average.

In a similar way to experiment 1, 416 interest terms extracted for Subject A were evaluated in terms of their appropriateness. (1) 384 were appropriate (92.3%), (2) 28 were appropriate if edited (6.7%), and (3) four were inappropriate (1.0%). These rate are quite similar as to those in Experiment 1. They show that the algorithm for extracting interest terms works well.

The subjects were asked whether terms from link texts or from documents matched their interest terms. All subjects said that link texts were more helpful. Table 1 lists the evaluations of the interest terms. Interest terms occurring more than twice were evaluated by each subject. Regarding link texts, the average score was 3.79; regarding documents, it was 3.30. These results mean that extracting interest terms from link texts is better than from documents - in particular, newspaper articles - terms of the precision rate.

Table 2 lists the evaluations of the interest-space browser. The results show that the system

Table 2: Evaluation of an interest space browser.

Question		A	B	C	D	Mean
Test 1	Q1	4	4	4	5	4.25
	Q2	4	1	4	5	3.50
	Q3	3	3	5	5	4.00
	Q4	4	4	4	3	3.75
	Q5	4	5	4	4	4.25
Test 2	Q1	3	4	4	5	4.00
	Q2	2	1	3	4	2.50
	Q3	2	2	4	5	3.25
	Q4	3	4	3	3	3.25
	Q5	4	5	4	4	4.25

successfully displays the users' interest space (averages: 4.25 in Test 1 and 4.00 in Test 2), and is helpful for recalling the user's past (averages: 4.25 and 4.25). The evaluations in Test 1 are better than those in Test 2. We consider that this is because it is more difficult to see the screen when terms and icons overlap, and the similarity of terms and icons is sometimes confusing. Therefore, the displaying algorithms and user interface must be improved.

Overall, the results suggest that the algorithm of extracting interest terms is useful, the system can appropriately displays the user's interest space, and the system is useful for recalling the user's past.

6 Related Work

Forget-me-not[4] is a system that reorganizes information like that in human episodic memory. The significance of the research was to present a new issue and a model which supports human memory. However, the system was developed before the age of the Internet and there were few facilities concerning the Web.

WebWatcher[5] and Letizia[6] learn user's interests for navigating as they navigate the Web. In contrast, our primary aim is to help users construct externalized memory. Web Forager[7] and Data Mountain[8] visualize a bookmark. The thinking-space browser, however, does not only deal with bookmarks and web pages but also memoranda of the users. MosaicG[9] and PadPrints[10] visualize a user's history using a tree structure. In contrast, the interest-space browser clusters and displays keywords contained in the history.

Much research[11] and practical systems (e.g., "Inspiration") for supporting creative activities relates to our work. The main difference between such systems and ours is the concept of the research. That is, the externalized memory is a broader concept by which creative thinking is per-

formed as a part of the control processes.

7 Conclusions

We have developed a system called Memory-Organizer that helps users to construct "externalized memory" and supports their creative activities. The system consists of (a) a thinking-space browser, which helps users to input and edit externalized memory and support creative thinking; (b) an overlay web browser, which helps users to gather, clip, and create externalized memory while browsing the Web; and (c) an interest-space browser, which helps users to recall externalized memory by arranging it on the user's interest space in chronological order. We used Memory-Organizer for information acquisition and reorganization while browsing the Web. It's performance was evaluated in two experiments, which revealed that (1) the algorithm of extracting interest terms from web pages works well and that (2) the interest-space browser can successfully display the user's interest space, and the user can recall their past by using this browser and reorganize previously browsed web pages of newspaper articles accordingly.

References

[1] Maeda, H., Hirata, T., and Nishida, T.: CoMeMo: Constructing and Sharing Everyday Memory, Proceedings of the Ninth International Conference on Tools with Artificial Intelligence (ICTAI'97), pp. 23-30 (1997).

[2] Murakami, H., Taki, R., Takashiro, T., and Nishida, T.: Chapter 5, Associative Representation for Personal Memory Management, in Nishida, T. (Ed.), Dynamic Knowledge Interaction, pp.131-181, CRC Press (2001).

[3] Kinoshita, E.: Wakariyasui-Suugaku-Model-Ni-Yoru-Tahenryokaiseki-Nyumon (in Japanese), Keigakushuppan (1987).

[4] Lamming, M., and Flynn, M.: "Forget-me-not" Intimate Computing in Support of Human Memory, XRCE Technical Report: EPC-1994-103 (1994).

[5] Armstrong, R., Freitag, D., Joachims, T., and Mitchell, T.: WebWatcher: A Learning Apprentice for the World Wide Web, Procddings of AAAI Symposium on Information Gathering from Distributed, Heterogeneous Environments, pp.6-12 (1995).

[6] Lieberman, H.: Letizia: An Agent That Assists Web Browsing, Proceedings of IJCAI95, pp.924-929 (1995).

[7] Card, S. K., Robertson, G. G., and York, W.: The Web Book and the Web Forager: An Information Workspace for the World-Wide-Web, in Proceedings of ACM CHI'96, pp.111-117 (1996).

[8] Robertson, G., Czerwinski, M., Larson, K., Robbins, D. C., Thiel, D., and Dantzich, van M.: Data Mountain: Using Spatial Memory for Document Management, in Proceedings of ACM UIST'98, pp.153-162 (1998).

[9] Ayers, E. Z., and Staski, J. T.: Using Graphic History in Browsing the World Wide Web, in Proceedings of WWW4 (1996).

[10] Hightower, R. R., Ring. L. T., Helfman, J. L., Bederson, B. B., and Hollan, J. D.: Graphical Multiscale Web Histories: A Studey of Padprints, in Proceedings of ACM Hypertext'98, pp.58-65 (1998).

[11] Young, L. F.: Chapter 8, Idea Processing Support: Definitions and Concepts, In Decision Support and Idea Processing Systems, Wm. C. Brown Publishers, pp.243-268 (1988).