Development of an Information-classification System Based on Access Logs

Noriyuki Hayashi†*, Keita Ooi, and Takeya Mukaigaito*

Abstract

With the amount of information on the Internet continuing to increase rapidly as broadband and mobile communication environments expand, information-classification technology is essential for retrieving desired information efficiently from the wealth of information on the Internet. NTT Information Sharing Platform Laboratories is developing an information-classification system based on access logs. We have developed a new Web browsing system called C-Chart using this information-classification technology. This enables Web browsing using information classification by C-Chart.

1. Need for information classification

In recent years, we have seen the rapid expansion of the broadband environment through FTTH (fiber to the home) and ADSL (asymmetric digital subscriber line) and of the mobile environment through cellular phones equipped with i-mode and other Internet-access systems. This development has helped produce a flood of information on the Internet, making it even more difficult for users to access the information they need.

To find information on the Internet, Web users use search services like Google and Yahoo! or goo in Japan [1]. In services of this type, the user enters keywords related to the desired information and then attempts to locate that information from among the limited number of search results returned. However, the amount of information on the Internet continues to increase, and in the future, we can expect the number of search results to be huge. Thus, finding desired information from all the results presented may take considerable time, and one could fail to find the desired information even after spending all that time (**Fig. 1(a)**).

Some Web sites, especially those involved with e-commerce, customize and present information for each user as part of their services. Customization of this kind, however, faces potential problems. For example, as the number of users and amount of information increase, the processing required to customize information for each user will likewise increase and it may become impossible to display customized information in real time. In addition, realtime customization of information may even degrade service quality and prevent optimal customization from being achieved. In short, customization of user information may make it difficult to provide the service itself (Fig. 1(b)).

NTT Information Sharing Platform Laboratories advocates information classification as an effective means of solving the above issues. This is because classifying the information in a huge number of search results can make the information-selection process more efficient. Information classification can also benefit services that customize information for each user. It can be used to classify users having similar interests, preferences, and behaviors into the same category so that they can be processed together as a group. This will eliminate the need for individual user processing and reduce the computational load.

2. Information-classification technology based on access logs

NTT Information Sharing Platform Laboratories is currently researching and developing information-classification technology based on access logs [2]. In this approach, some kind of relationship is considered

† NTT Information Sharing Platform Laboratories
Musashino-shi, 180-8585 Japan
E-mail: hayashi.noriyuki@lab.ntt.co.jp

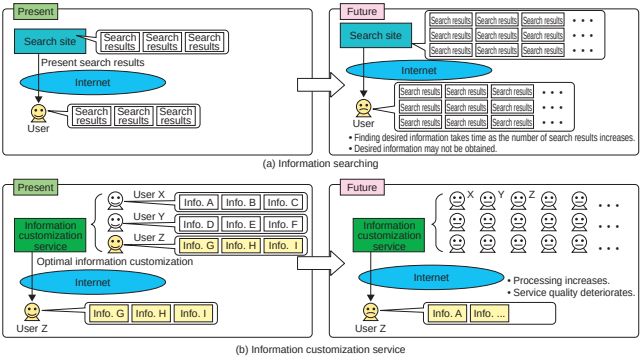

Fig. 1. Future issues in information searching.

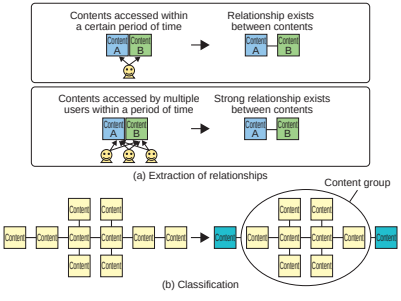

Fig. 2. Information-classification technology based on access logs.

to exist among contents accessed by the same user within a certain period of time. Furthermore, a strong relationship is considered to exist among certain contents that have been accessed by multiple users within the same period of time. Using access logs in this way enables relationships between contents to be derived (**Fig. 2**). In particular, contents that are found to be related are connected, and if the number of interconnections exceeds a certain value, those contents are categorized as a content group having a strong relationship. In this technology, the system updates the classifications in the neighborhood of accessed content every time an access log is updated. Limiting the range of computation in this way reduces the computational load when classifying huge amounts of information and enables realtime information classification to be performed.

3. C-Chart

We have developed a new Web browsing system called C-Chart[*1] using this information-classification technology [3]. C-Chart does more than simply provide links. It also allows users to browse through information using information classification based on access logs. The provision of such categorized information can make information

*1 C of C-Chart expresses cluster, context, and content.

browsing more efficient. It should also enable users to make discoveries that were not possible with previous browsing systems. Its configuration is shown in **Fig. 3**. The server consists of a proxy that obtains access logs of user information browsing, an engine that classifies information based on the obtained access logs and generates classification data, and a database that stores that classification data.

The system converts this classification data into metadata in RDF*2 format and then sends the data to a user terminal. This process gives generated data general-purpose features so that it can be applied to a variety of Web services and applications and not just the client application presented here. For this reason, we expect C-Chart to find widespread use.

The user terminal in this system consists of Internet Explorer and the C-Chart client. The client links with the browser to display information-classification results from Web browsing in the client's window (**Fig. 4**). The server analyzes Web access from Internet Explorer in real time and reflects the analysis results on the C-Chart client. The nodes shown on the client represent URLs so that clicking on a node displays the Web page corresponding to that URL on the browser. Since a click is reflected in Internet Explorer in the node in the C-Chart client, browsing by clicking nodes is also possible. Since browsing can also be carried out from an information classification result in addition to browsing by clicking a conventional Web link, even more efficient browsing is possible.

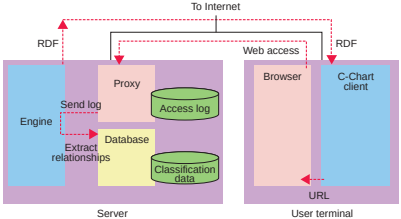

Fig. 3. C-Chart system configuration.

*2 RDF (resource description framework): A standard for exchanging metadata on the Web. It prescribes a description method by the World Wide Web Consortium (W3C) and a method of exchange between computers.

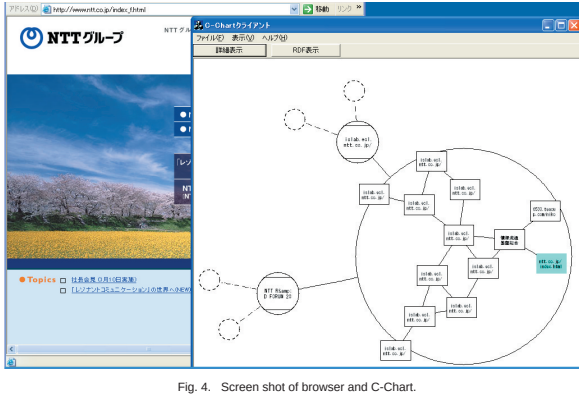

Fig. 4. Screen shot of browser and C-Chart.

Services that classify search results for presentation are being provided by various companies including Vivisimo [4], Grokker2 [5], and TouchGraph [6]. These services, however, classify search results while C-Chart classifies and displays items related to search results. This means that C-Chart can be appropriately combined with technologies that classify search results.

4. Expansion to various services

The C-Chart system targets URLs in access logs and classifies that information. The underlying technology, however, can analyze anything obtained from access logs. When the access log of a search site, for example, is used, the keywords entered by users for searching can be analyzed to extract relationships among them and to classify that information accordingly.

Moreover, the purchasing log of an e-commerce site can be used to extract relationships among purchased products to classify purchasing information and present users with recommended products. Although conventional recommendation services often present similar products individually as recommended items (for example, another book by the same author), our technology can classify offered products so that an e-commerce operator can present recommended products in a more efficient manner.

5. Application to the Semantic Web

The Semantic Web*3 has been receiving much attention as technology for increasing the accuracy of search results and making it easier for users to utilize the flood of information on the Web. Before the Semantic Web can be created, however, many issues must be resolved, with one in particular being the generation of metadata [7]. Considering that our technology can be used to extract information structures from access logs and generate and assign metadata, we are also studying its application to the Semantic Web.

6. Future plans

This article described information-classification technology based on access logs now being researched and developed in NTT Information Sharing Platform Laboratories and introduced the new C-Chart browsing system that employs this information-classification technology. We plan to improve its performance and expand it to services and fields that make use of such characteristics.

*3 A technology for giving Web information that describes content (metadata) so that computers can understand content and automatically process it. The idea here is to collect Web metadata into a database and to extract and structure the knowledge to make the Web more useful. The Semantic Web aims to provide more intelligent services. If this technology becomes practical, the search results will become more accurate and users will find it easier to use the massive amount of information on the Web.

References

[1] http://www.goo.ne.jp/
[2] K. Ooi, N. Hayashi, and T. Mukaigaito, "C-Chart," Semantic Web Conference 2003, Proceedings, 6-6, Tokyo, Japan, Nov. 2003 (in Japanese).
[3] N. Hayashi, K. Ooi, and T. Mukaigaito, "Related Information Presentation System using Access Log," 2004 IEICE General Conference, B-16-29, Tokyo, Japan, Mar. 2004 (in Japanese).
[4] http://vivisimo.com/
[5] http://www.groxis.com/
[6] http://www.touchgraph.com/
[7] INTAP, "Annual Report for Fiscal Year 2002 of the Semantic Web Committee," (in Japanese).

Noriyuki Hayashi
Research Engineer, Software Architecture Project, NTT Information Sharing Platform Laboratories.
He received the B.E. and M.E. degrees in electronics engineering from Hokkaido University, Sapporo, Hokkaido in 1991 and 1993, respectively. He joined NTT in 1993. He has been engaged in R&D of context oriented sharing technology. His current research interests include context-awareness and open source systems. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE).

Keita Ooi
Engineer, Software Architecture Project, NTT Information Sharing Platform Laboratories.
He received the B.S. and M.S. degrees in mathematical science from Kyoto University, Kyoto in 1996 and 1999, respectively. In 1999, he joined NTT Information Sharing Platform Laboratories, Tokyo. His current interest is collaborative filtering.

Takeya Mukaigaito
Senior Research Engineer, Software Architecture Project, NTT Information Sharing Platform Laboratories.
He received the B.E. degree in engineering science from the University of Tsukuba, Ibaraki in 1989. In 1989, he joined NTT Communication and Information Processing Laboratories, Kanagawa. His current interest is context oriented application architecture.