

## Information Navigation Techniques Based on the Bird's-eye View Approach

*Hidenori Okuda<sup>†</sup>, Harumi Kawashima, Yoshihide Sato, Shinji Miyahara, and Toru Sadakata*

### Abstract

In next-generation portal sites, one approach to aiding Internet users who do not have clear search objectives is to present a bird's-eye view data map that arouses the user's interest in a topic. This article describes a technique for extracting topic words from blogs (Weblogs) or Web news sites and interface techniques for displaying a large number of keywords and contents on a limited-size screen.

### 1. Web navigation methods

According to the Institute for Information and Communications Policy of the Ministry of Internal Affairs and Communications [1], in February 2004 there were 85.9 million Web pages in the jp domain alone on the Internet. Two methods of navigating this enormous body of information are the search-based approach and the bird's-eye view approach (**Fig. 1**).

The search-based approach is the most efficient for users who have a clear search objective: they can search for the information they desire by using directory services and search services. They can also get easy access to updates in Website contents by using the push-type service RSS (rich site summary, RDF (resource description framework) site summary, or really simple syndication). Users who do not have a clear search objective, however, cannot use search services effectively if they cannot think of keywords to enter to aid in their search. And they cannot use RSS because they cannot specify the Websites from which they wish to receive updated infor-

mation. In such cases, the bird's-eye view approach is effective because it allows users to see displays of overall data objects on the Internet, in the form of bird's-eye view maps and from them find information that attracts their interest to guide them in their search. In this way, the bird's-eye view approach serves as an additional tool for navigating Web information, supplementing established search services. This has made research and development on the bird's-eye view approach a subject of major importance.

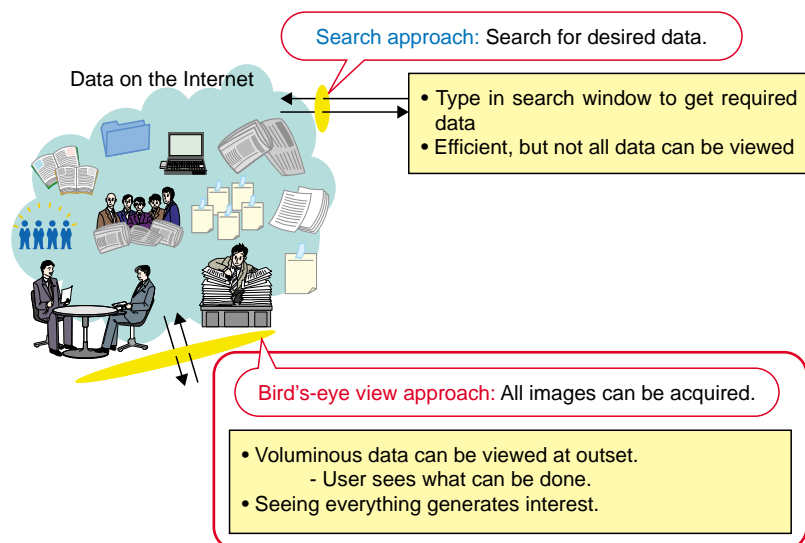


Fig. 1. Methods of navigating information on the Internet.

<sup>†</sup> NTT Cyber Solutions Laboratories  
Yokosuka-shi, 239-0847 Japan  
E-mail: okuda.hidenori@lab.ntt.co.jp

## 2. Latest topic word extraction technique

Various methods are available for showing bird’s-eye views of information on the Internet, but in this paper we focus on the “latest topic word extraction” technique [2], which enables users to see the degree to which certain information stands out from the rest. This technique uses blogs (Weblogs) and Web news sites to show outstanding information and extracts keywords that directly express contents relevant to fresh topics. The user can ascertain which topics are particularly prominent on the Internet at a given time merely by observing these keywords.

Since Web news sites carry articles on news stories that are of interest to large segments of the public, users can judge which stories are most prominent at that time by observing those for which many articles have been compiled from multiple news sites at the same time. In the latest topic word extraction technique, articles are first broadly sorted by genre, e.g., business, science, or sport. They are then grouped into specific topics by clustering them on the basis of words and expressions appearing in them. Representative words and expressions in each cluster are then extracted as keywords for a given topic, i.e., topic words. The flow of the extraction of the latest topic words from Web news sites is shown in Fig. 2.

With blogs, on the other hand, people write in diary form about aspects of topics that stir their emotions, and Web users can use blogs to ascertain which topics are stirring passions around the globe by noting those that are treated in many different blogs. Blogs

have certain unique characteristics: one is that bloggers often include multiple topics in one posting; another is that different bloggers may use different words and expressions to refer to the same thing in talking about the same topic. In dealing with blogs, therefore, clustering based on the mere appearance of words and expressions cannot be applied to the latest topic word extraction technique in its current form. Instead, topic words are extracted by putting weight on words and expressions that have begun to appear very frequently in recent blogs.

The system layout of the latest topic word extraction technique is shown in Fig. 3. The system’s article crawler/classifier collects news articles and blog postings from the Internet and sorts them by genre. The topic word extractor extracts topic words in the manner described above. The topic word integrator integrates topic words of two systems and hands them over to the client as data output.

## 3. Keyword display interface

Here, we introduce the display interface HotWindow+ as an interface technology for efficiently displaying on a limited-size screen a large number of keywords extracted from information on the Internet. The HotWindow+ system displays keywords extracted with the latest topic word extraction technique, i.e., topic words on the screen of a client (personal computer).

An example of a screen of the HotWindow+ display interface is shown in Fig. 4. The topic genres appear

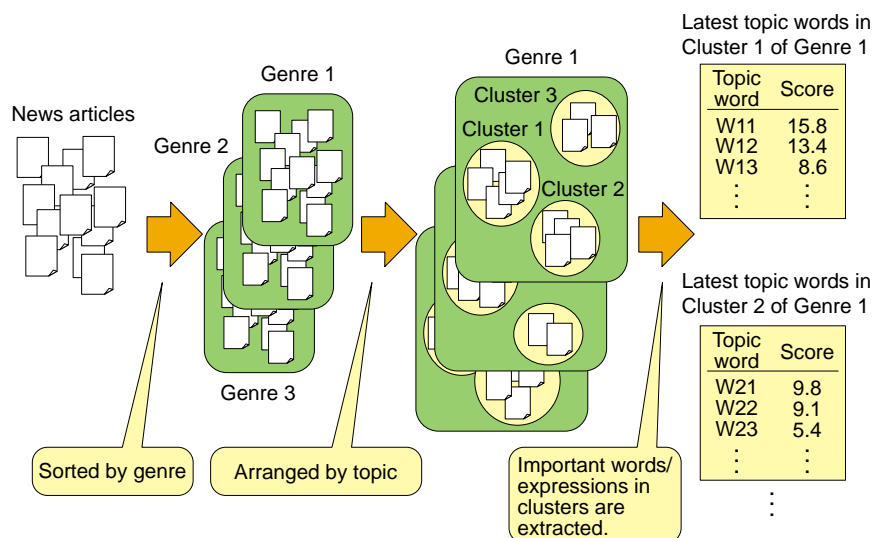


Fig. 2. Flow of the extraction of the latest topic words from Web news sites.

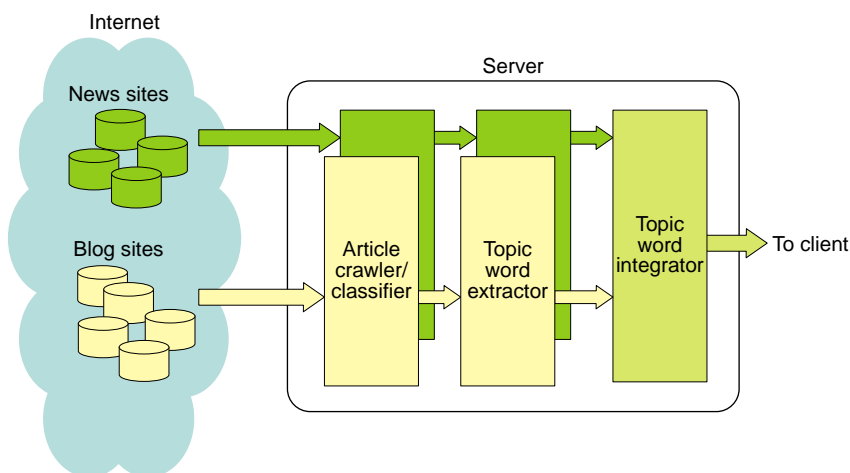


Fig. 3. System layout of the “latest topic word extraction” technique.

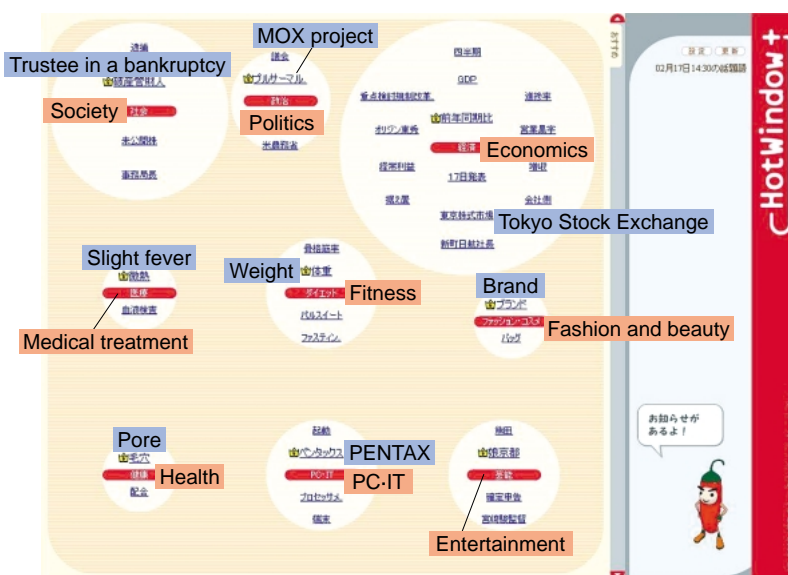


Fig. 4. Example screen of the HotWindow+ display interface.

in circular regions and topic words for each genre are displayed. When the user clicks the genre title in the center of a region, the region is magnified and a larger list of topic words appears. When this is done, the other circular regions around the clicked one do not disappear but remain displayed in small form, so the entire image of the information can still be seen.

Moreover, since it is inconvenient for users who do not have a clear search objective to have to actively manipulate the data they are perusing, the system automatically marks genres that have been magnified. Marking the magnified genres in this way effectively spotlights them, and switching the spotlight

from one genre to another makes it possible to call the user’s attention to keywords for a wide variety of genres.

#### 4. Interface for displaying content items numbering in the hundreds of thousands

Here, we introduce “High Choukan” [3] as an interface for displaying contents on the order of hundreds of thousands of items, where *choukan* means bird’s-eye view in Japanese. This interface effectively steers users who do not have a clear search objective towards contents that will likely be of interest to

them. It does so by giving them an impression of the voluminous contents, showing what types of contents are available and stimulating their interest in the contents.

To show users what sort of contents are available, High Choukan classifies contents by color according to information type and locates them under background images relevant to the content type. This enables users to get a bird's-eye view of the overall concept of the contents.

**Figure 5** shows an example screen, constructed using High Choukan, which offers a bird's-eye view of information about shops, stores, and other such establishments. The contents in the example comprise some 160,000 data items on dining establishments such as restaurants and hotel restaurants in the 23 wards of Tokyo. A map of the area serves as the background image. Dining establishments are grouped by type into colored circles and located over background images showing the establishment's address and location. Zooming and panning functions are available for easier navigation. If users wish to acquire details about particular areas they are interested in, they can do so by zooming in on the area in question. If they also wish to obtain information about adjoining areas, they can do so by moving horizontally over to them. If they are interested in finding out about more than one establishment in a given area, they can obtain a list of other establishments in the area to choose from by clicking on the background image.

An enormous amount of calculation is required to

arrange hundreds of thousands of items for bird's-eye viewing within the screen space and allow them to be moved around within the area. To reduce the content-display calculation required, High Choukan acquires and displays on the screen only the minimal background images necessary, i.e., those needed to synthesize into image form the contents and the background space where they are allocated. This greatly reduces the amount of calculation needed for zoom and pan operations, even when the number of contents is extremely large. It also enables contents to be moved and displayed while converting synthesized background images into various different scaled-down forms. **Figure 6** shows the layered structure of background images, in which the images are structured in upper and lower scaled-down forms. From among these images, the ones necessary for the user's position in the space are selected and displayed.

## 5. Future developments

We intend to research methods of clustering not only news articles but also blog articles to enhance the targetability of topic words extracted using the latest topic word extraction technique. Since blog topics are extremely useful in corporate marketing activities, we will also attempt to expand the usage of blog topic correlation and various other analysis parameters. As for display interfaces, we will strive to develop interfaces for displaying blogs, message boards, and other contents showing a significant inter-relationship among topics.

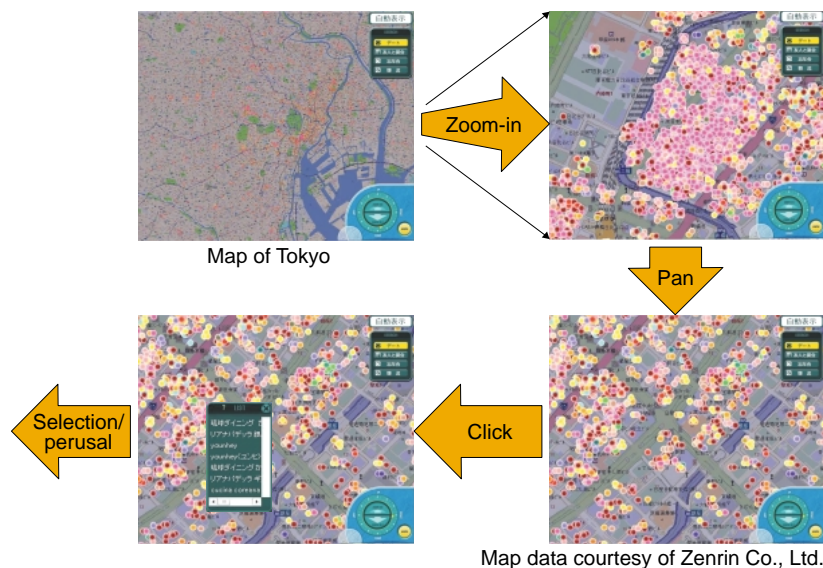


Fig. 5. Example screen constructed using High Choukan.

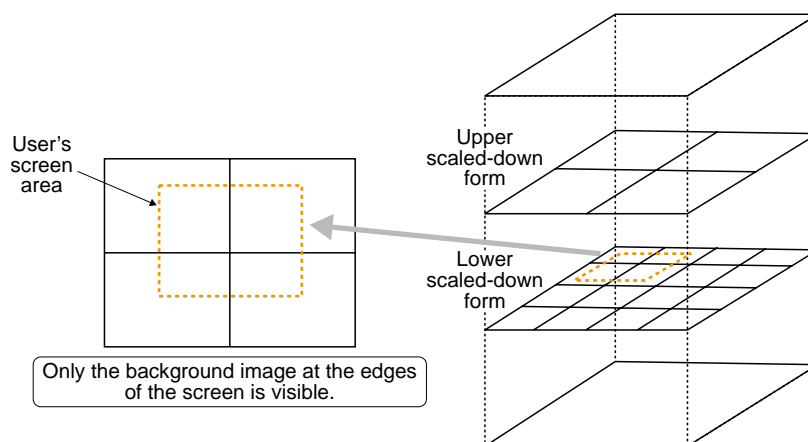


Fig. 6. Layered structure of background images.

## References

- [1] <http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2006/2006-1-01-2.pdf> (in Japanese).
- [2] Y. Sato, H. Kawashima, T. Sasaki, and M. Oku, "Latest Topic Words Detection from Chronological News Stream," IEICE Technical Report, 2005-NL-168, pp. 1-6, 2005 (in Japanese).
- [3] S. Miyahara, T. Sadakata, and M. Oku, "Navigation Interface that Uses Prosperity of Other Users Intended for Several 100,000 Contents," IEICE Technical Report, 2005-HIP-49, pp. 159-162, 2005 (in Japanese).



### Hidenori Okuda

Senior Research Engineer, Supervisor, Media Computing Project, NTT Cyber Solutions Laboratories.

He received the B.E. and M.E. degrees in precision machinery engineering from the University of Tokyo, Tokyo, in 1986 and 1988, respectively. He also received the M.S. degree in computer science from Stanford University in 1994. In 1988, he joined NTT Human Interface Laboratories and mainly engaged in R&D of visual communication systems. From 1997 to 2003, he engaged in multimedia business development in the business sections of NTT and NTT-ME. He is currently interested in techniques for mining information from voluminous consumer-generated media. He is a member of the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan and the Institute of Image Information and Television Engineers.



### Harumi Kawashima

Senior Research Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.

She received the B.E. and M.E. degrees in computer science engineering from Yamanashi University, Yamanashi, in 1988 and 1990, respectively. Since joining NTT Human Interface Laboratories in 1990, she has mainly been engaged in R&D of communication and information systems. She is a member of IEICE.



### Yoshihide Sato

Research Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.

He received the B.E. degree in electrical and electronics engineering and the M.E. degree in informatics from Kyoto University, Kyoto, in 2000 and 2002, respectively. He joined NTT in 2002. His current research interests include blog mining, document clustering, and keywords identification. He developed the extractor of latest topic words in the HotWindow+ system.



### Shinji Miyahara

Research Engineer, Innovative IP Architecture Center, NTT Communications.

He received the B.E. degree in electrical, computer, and systems engineering from Hiroshima University and the M.E. degree in information science engineering from Osaka University, Osaka, in 1998 and 2000, respectively. He joined NTT Cyber Solutions Laboratories in 2000 and was mainly engaged in interface research on the Internet. After these R&D activities, he moved to the Innovative IP Architecture Center of NTT Communications in 2006.



### Toru Sadakata

Senior Research Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.

He received the B.E. and M.E. degrees in applied physics from Waseda University, Tokyo, in 1989 and 1991, respectively. In 1991, he joined NTT Human Interface Laboratories, Tokyo. He was engaged in R&D of video handling and human interfaces. From 1997 to 2000, he worked in NTT Printec Corporation (now NTT Quaris), Tokyo. He was engaged in the development of a color printing system. In 2000, he moved to NTT Cyber Solutions Laboratories. His current research interests are contents handling and human interfaces. He is a member of IEICE, the Human Interface Society of Japan, and the Association for Computing Machinery.