Navigational Information Retrieval Technologies that Help Users Reach the Desired Information

Ryoji Kataoka†, Hiroyuki Toda, Yukio Uematsu, Ko Fujimura, Katsuji Bessho, and Shuichi Nishioka

Abstract

We present some next-generation information retrieval technologies that help users reach the information that they are seeking according to their situation and time.

1. Toward the era of navigational search service

With the proliferation of broadband connections, we are using the Internet more often in a variety of ways in daily life. In recent years, the rapid spread of search services has been especially remarkable. Truly effective search functions are essential for various portal services handling a huge volume of information because the search function is the main service that users use among the various ones provided by portals. For example, there are search services designed for multimedia such as images, movies, and music on the web and ones designated for news and blog (Weblog) articles that are being updated continuously.

Search services continue to evolve. The first-generation search service, a directory-type search service, is a major information retrieval tool and has grown along with the Internet. The second-generation search service is a search-engine-type service based on direct searching by keywords and it has grown with the volume of information available. In an age in which the way people use the Internet has diversified and broadened, NTT Cyber Solutions Laboratories thinks that the third-generation search service type will be a navigational search service that assists users in encountering the information they need in their current situation. In this article, we present some technologies that will enable the next-generation navigational search service by referring to some of our research.

† NTT Cyber Solutions Laboratories
Yokosuka-shi, 239-0847 Japan
https://www.ntt.co.jp/cclab/e/contact/index.html

2. Research topics

2.1 TopicMaster: a news article retrieving and clustering system

With the amount of information available on the Internet continuing to increase day by day, the amount of information produced by a search is likely to continue to increase. Most existing search services can rank search results based on certain criteria, but it remains extremely difficult to find the desired information from the enormous list of search results. Also, it is all too common for the user to become irritated when he/she cannot come up with another appropriate keyword to prune the search results.

To alleviate this weakness of existing search services, we have invented a technique that actively extracts proper names (such as the names of people, organizations, and places) included in the search results as topics and generates and displays a category structure (topic tree) that properly categorizes the search results based on the extracted topics (**Fig. 1**) [1]. We applied this technique to a news service to develop TopicMaster, a system that retrieves and categorizes news articles. TopicMaster enables users to overview an enormous number of search results through the topic tree and easily prune them by clicking on a topic in the tree.

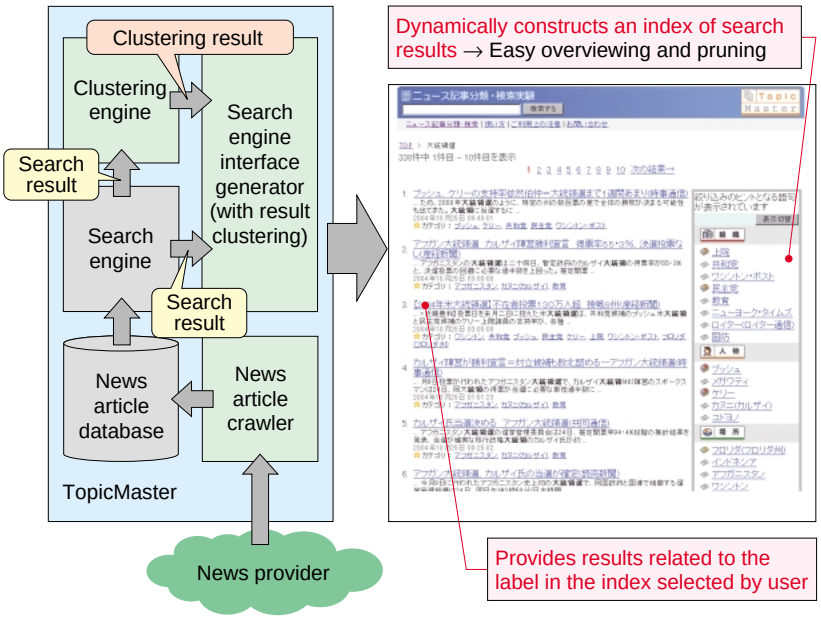

Fig. 1. TopicMaster.

2.2 MultiMedia Meister: efficient search system for massive multimedia contents on the web

Regardless of whether the user wants a still image or a movie, it is inconvenient and difficult to find exactly the desired image from a huge number of search results. Therefore, especially for multimedia content searches, portal services must offer a search function for finding images that are difficult to describe in words. The traditional search method uses keywords, but it is difficult for the user to determine which keywords will locate the desired image. For example, when a user wants a particular image of Mt. Fuji, it will be extremely difficult to find it unless the uniqueness of the image can be well captured by keywords known to the user. NTT Cyber Solutions Laboratories has improved the function of its original fast full-text search engine to create MultiMedia Meister, a search system that provides a website search function suitable for multimedia [2].

MultiMedia Meister can narrow down the search results by classification or visual similarity according to the features of a reference image or a movie (color and shape) in addition to the traditional keyword search method (**Fig. 2**) [3]. With MultiMedia Meister, the user not only surveys the images that were retrieved, but can also select a returned image and use it to refine the search process in an intuitive manner.

It normally takes time to analyze the similarity of the images based on their features, but MultiMedia Meister can search billions of images at a practical speed because it can simultaneously retrieve text information and image features at high speed.

2.3 BLOGRANGER: multifaceted blog search system

Information distributed by blog services contains what consumers really think. For example, blogs contain information and recent topics that have not been taken up by the mass media and present opinions about products or services. Such information is expected to be used for marketing, and a number of blog search services have appeared. However, since most existing services show the search results ranked just by date or by blog access frequency, users sometimes have difficulty in reaching the information they want in blogs, because of the massive amounts of information and the frequent updating.

We have developed a blog search system called BLOGRANGER to overcome the weaknesses of existing services. It shows search results categorized and arranged by keywords in a multifaceted manner through four useful kinds of filters. This enables users to search for information not only by keywords but also by four viewpoints: topics (categorization by proper name just like TopicMaster), bloggers (authors of blogs), link destinations (Web pages referenced by blog articles), and sentiments (evaluations or impressions such as “interesting” and “wonderful”). The search results can be categorized and arranged by these four viewpoints (**Fig. 3**).

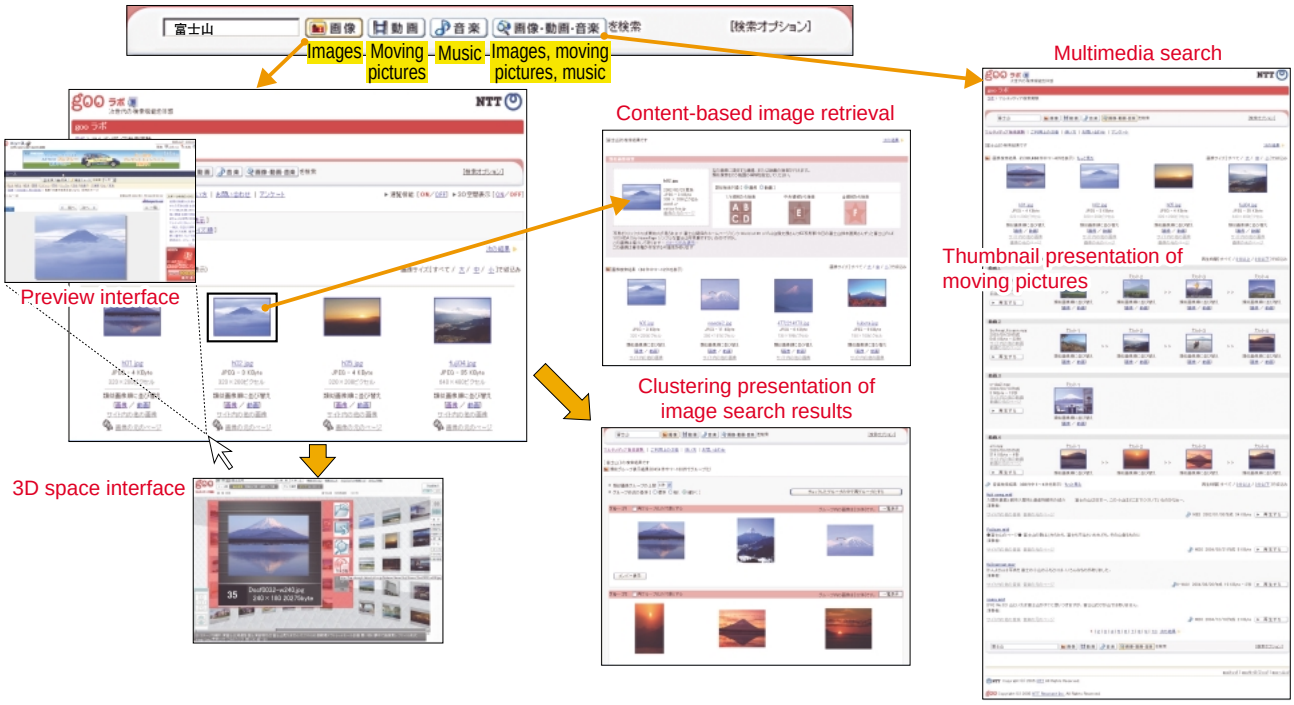

Fig. 2. Multimedia Meister.

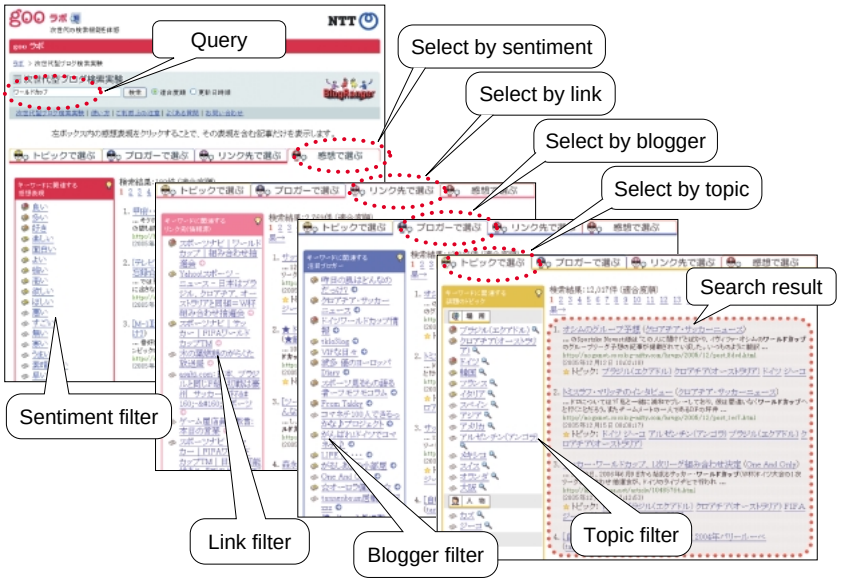

Fig. 3. BLOGRANGER.

Out of these four filters, the blogger filter makes use of one of the characteristics of blogs: the articles are edited by one author (blogger). To see if the blogger writes good articles to which many people want to link or if the blogger links good articles to his/her blog, BLOGRANGER numerically evaluates blog articles based on their link relationships. It then ranks blog articles by popularity [4]. By getting a list of bloggers who have written highly ranked articles, users can easily identify useful articles written by reliable bloggers from among the enormous number of search results.

2.4 Kanshinji Antenna: Japanese-language concept search system

The technologies we have presented here are navigation-oriented and assist the user by narrowing down the search results: they focus on the problem that search services generally return a huge number of search results to the user. Another problem with existing search services is that search results are limited by the information contained in the input keyword, so they may contain very little in the way of desired information if the first keyword is inappropriate. Therefore, the next-generation search services need navigation technologies that can expand the search target according to the user's search goal.

To solve those problems, we are developing a Japanese-language concept search technology. This determines the concept of the words or phrases input by the user. For searching, it collects documents based not only on the input keyword but also on texts similar in concept to the keyword or phrase input by the user.

There is one more feature. Since concept search services require a dictionary to determine the similarity of semantic content between words, new words that are not registered in dictionaries, because they appear on the Internet for the first time, degrade search precision. Our technology solves this problem. If an unknown word is detected in the search results, the semantic content of the unknown word is automatically identified and registered based on the similarity of semantic content between words in the search results and those already registered in the dictionary [5]. Thus, this technology enables concept searches to be provided by Internet search services for the first time.

We have developed Kanshinji Antenna, a system that allows Japanese-language concept searching to be added to the latest news and blog articles search service (**Fig. 4**). When the user registers a keyword or sentence related to the information they want as an "antenna", they can retrieve all information associated with their interest.

2.5 TopicAlert: push-type topic notification system

In order to discover information on service-providing sites such as news websites or blogs that frequently release the latest information, it is important that these sites provide users with the latest information on a timely basis. This has led to the gradual proliferation of alert services that send notices to users by e-mail or other means when the desired information is released to the public.

Since existing alert services allow users to set only simple search conditions, it is difficult to receive just the information that suits the user's taste from blogs.

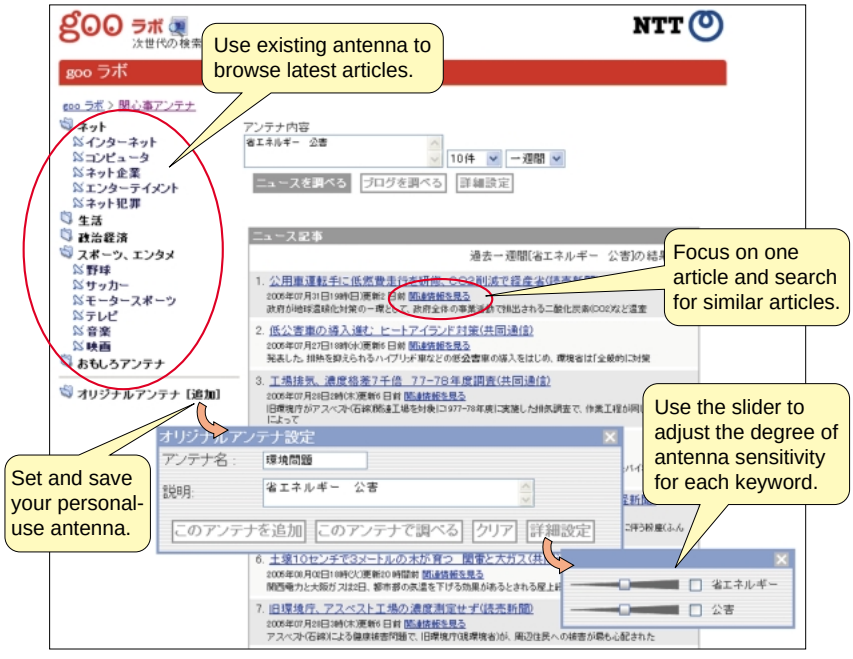

Fig. 4. Kanshinji Antenna.

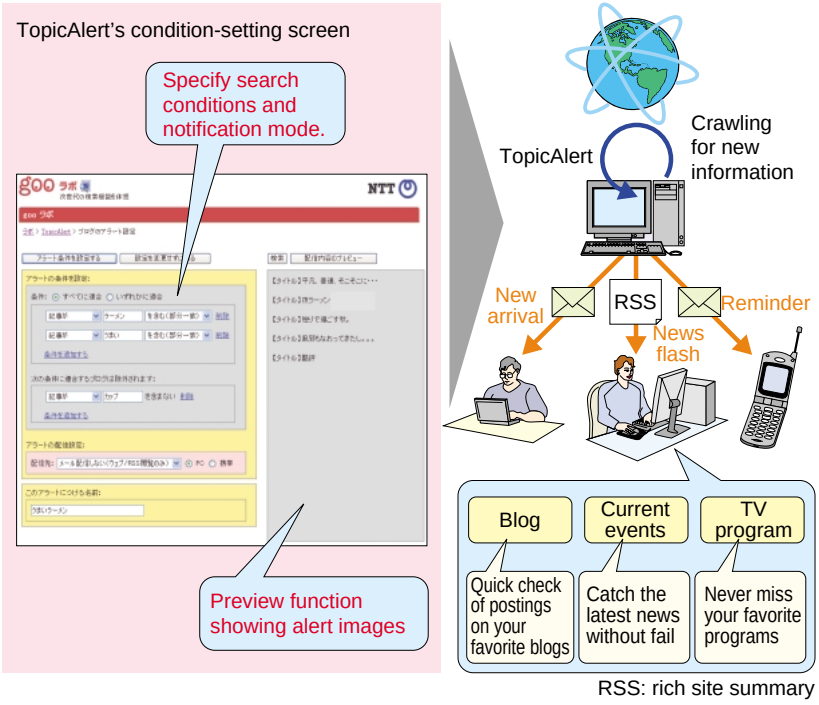

Fig. 5. TopicAlert.

Moreover, since the existing alert services can match the search conditions users already set with newly arriving information only in a one-by-one manner, they demand extensive hardware resources to provide immediate services, which leads to high operating costs.

Our push-type topic notification system TopicAlert uses high-speed XML (extensible markup language) filtering that enables efficient information matching to achieve quick selection of information by setting specific search conditions while keeping the traditional system environment (**Fig. 5**) [6]. Here, we explain how this system works. The XML filtering method converts and structures the search conditions of topics that the user wants into the XPath[*1] format. With this technology, the system compares those topics with the search criteria such as keywords or genres input by other users and integrates the overlapping search criteria. In this way, the system only has to search using the integrated search criteria once. XML filtering thus makes the matching process more efficient. New information extracted from the Internet is converted into XML format data. TopicAlert reads this XML format data to match the integrated search criteria from the top. In this way, TopicAlert can identify all users who need the information while minimizing the matching process because it can narrow down the search criteria that match newly arriving information by using the integrated search criteria.

*1 A standard that defines the description method designating the specific elements in an XML document.

3. Future developments

We have verified the technologies presented in this article in terms of their technological and operational usefulness through tests on "goo Labs" [7], the experimental site built into NTT Group's integral portal site "goo". MultiMedia Meister has been commercialized as a search engine for a multimedia search service for images, moving pictures, and music on goo. We are committed to continuously developing the next-generation information search technologies to provide users with safe, reliable, and convenient search services.

References

[1] H. Toda and R. Kataoka, "A Clustering Method for News Articles Retrieval System," in Poster Proceedings of the WWW 2005 Conference, 2005.
[2] H. Takeno and T. Inoue, "Distributed information gathering and full text search system Infobee/Evangelist," NTT R&D, Vol. 52, No. 2, 2004 (in Japanese).

[3] Y. Uematsu, R. Kataoka, and H. Takeno, "Clustering Presentation of Web Image Retrieval Results using Textual Information and Image Features," in Proceedings of the EuroIMSA 2006 Conference, 2006.
[4] K. Fujimura, T. Inoue, and M. Sugizaki, "The EigenRumor Algorithm for Ranking Blogs," in Proceedings of the WWW 2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics, 2005.
[5] K. Bessho, O. Furuse, and R. Kataoka, "Concept Vector Generation Method Based on Co-occurrences between Words and Semantic Attributes," in Proceedings of the 20th Annual Conference of the Japanese Society for Artificial Intelligence, 2006 (in Japanese).
[6] M. Onizuka, F. Y. Chen, R. Michigami, and T. Honishi, "Incremental Maintenance for Materialized XPath/XSLT Views," in Proceedings of the WWW 2005, 2005.
[7] http://labs.goo.ne.jp/

Ryoji Kataoka
Senior Research Engineer, Supervisor, Media Computing Project, NTT Cyber Solutions Laboratories.
He received the B.E. and M.E. degrees in electronic engineering from Chiba University, Chiba, in 1985 and 1987, respectively. He joined NTT Communications and Information Processing Laboratories in 1987 and engaged in research on transaction processing and multimedia databases. He is currently engaged in R&D of information retrieval technologies. He is a member of the Information Processing Society of Japan (IPSJ).

Hiroyuki Toda
Research Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.
He received the B.E. and M.E. degrees in material science from Nagoya University, Aichi, in 1997 and 1999, respectively. He joined NTT Cyber Space Laboratories in 1999. Since then, he has been engaged in R&D of information retrieval, text mining, and clustering. He is currently a Ph.D. student at the University of Tsukuba, Ibaragi. He is a member of ACM SIGIR, IPSJ, and the Database Society of Japan (DBSJ).

Yukio Uematsu
Research Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.
He received the B.E. and M.E. degrees in computer science from Tokyo University of Science, Tokyo, in 2001 and 2003, respectively. He joined NTT Cyber Solutions Laboratories in 2003. Since then, he has been engaged in R&D of information retrieval, text mining, and web image retrieval. He is currently a Ph.D. student at Tokyo University of Science, Tokyo. He is a member of IPSJ and the Japanese Society for Artificial Intelligence.

Ko Fujimura
Senior Research Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.
He received the B.E. and M.E. degrees in electrical engineering and the Dr.Eng. degree in information engineering from Hokkaido University, Hokkaido, in 1984, 1986, and 1989, respectively. He joined NTT Information Processing Laboratories in 1989 and engaged in R&D of transaction processing systems and digital ticket systems and standardization of payment systems in the Internet Engineering Task Force and Infrared Data Association. Since 2003, he has been engaged in research on community mining and blog search engines. He is also a visiting professor of the University of Electro-communications, Tokyo. He is a member of IPSJ, the Institute of Electronics, Information and Communication Engineers (IEICE) of Japan, and the Japan Association for Social Informatics.

Katsuji Bessho
Research Engineer, Media Computing Project, NTT Cyber Solutions Laboratories.
He received the B.S. and M.S. degrees in mathematics from Osaka University, Osaka, in 1992 and 1994, respectively. He joined NTT in 1994. He is currently at the Media Computing Project, NTT Cyber Solution Laboratories and working on natural language processing technology. He is a member of IEICE, IPSJ, and the Association for Natural Language Processing.

Shuichi Nishioka
Research Engineer, Promotion Project 1, NTT Cyber Solutions Laboratories.
He received the B.E. degree in electrical and computer engineering and the Dr.Eng. degree in information media and environment sciences from Yokohama National University, Kanagawa, in 1995 and 2005, respectively. Since joining NTT Laboratories in 2005, he has been engaged in research on database management systems, copyright management systems, and XML processing systems. He is a member of IPSJ and DBSJ.