Proposal of Integrated Search Engine
of Web and TV Contents

Hisashi Miyamori[1]
Mitsuru Minakuchi[1]

Zoran Stejic[2]
Qiang Ma[1]

Tadashi Araki[2]
Katsumi Tanaka[1,3]

[1]Interactive Comm. Media and
Contents Group, NICT

3-5 Hikaridai, Seika-cho, Soraku-gun,
Kyoto, 619-0289 Japan
+81-774-98-6873

{miya,mmina,qiang}@nict.go.jp

[2]Software R&D Group, Ricoh Co., Ltd.
1-1-17 Koishikawa, Bunkyo-ku,
Tokyo, 112-0002 Japan
+81-3-6801-1467

{zoran.stejic,araki-t}
@nts.ricoh.co.jp

[3] Graduate School of Informatics,
Kyoto University
Yoshida-Honmachi, Sakyo-ku,
Kyoto, 606-8501 Japan
+81-75-753-5969

ktanaka@i.kyoto-u.ac.jp

ABSTRACT

A search engine that can handle TV programs and Web content in an integrated way is proposed. Conventional search engines have been able to handle Web content and/or data stored in a PC desktop as target information. In the future, however, the target information is expected to be stored in various places such as in hard-disk (HD)/DVD recorders, digital cameras, mobile devices, and even in real space as ubiquitous content, and a search engine that can search across such heterogeneous resources will become essential. Therefore, as a first step towards developing such next-generation search engine, a prototype search system for Web and TV programs is developed that performs integrated search of those content, and that allows chain search where related content can be accessed from each search result. The integrated search is achieved by generating integrated indices for Web and TV content based on vector space model and by computing similarity between the query and all the content described by the indices. The chain search of related content is done by computing similarity between the selected result and all other content based on the integrated indices. Also, the zoom-based display of the search results enables to control media transition and level of details of the contents to acquire information efficiently. In this paper, testing of a prototype of the integrated search engine validated the approach taken by the proposed method.

Categories and Subject Descriptors

H.2.4 [**DATABASE MANAGEMENT**]: Systems – *multimedia databases,* H.5.1 [**INFORMATION INTERFACES AND PRESENTATION**]: Multimedia Information Systems – *video.*

General Terms: Algorithms, Management, Documentation, Design.

Keywords: Search engine, information retrieval, information integration, Web content, TV programs, integrated search, chain search.

1. INTRODUCTION

Conventional search engines have been able to handle Web

content and/or data stored in PC desktop as target information. For example, there are search engines for specific media such as images on the Web, or for Web content with characteristics of rapid communication such as blogs. Desktop search can find various types of files in a PC by using full-text search. The information sources for search engines are gradually spreading from Web content to data stored in individual PCs.

In the future, however, the target information for search engines is expected to be stored in various places such as in hard-disk (HD)/DVD recorders, digital cameras, mobile devices, and even in real space as ubiquitous content and a search engine that can search across such heterogeneous resources will become essential. Since conventional search engines have been based on link analysis and full-text search, a new type of retrieval method will be necessary which is not dependent of these methods and which can rank and find various types of media stored in various places according to context in an integrated manner.

In this paper, as a first step towards developing such next-generation search engine, a prototype search system for Web and TV programs is developed that performs integrated search of those content, and that allows chain search where related content can be accessed from each search result. The integrated search is achieved by generating integrated indices for Web and TV content based on vector space model and by computing similarity between the query and all the content described by the indices. The chain search of related content is done by computing similarity between the selected result and all other content based on the integrated indices. The proposed method provides:

- Simultaneous search for Web and TV contents that match with the given keywords,

- Search for recorded TV programs that relate to the Web content being browsed

- Search for Web content or other TV programs that relate to the TV programs being watched.

In this paper, testing of a prototype of the integrated search engine validated the approach taken by the proposed method.

2. OVERVIEW OF PROPOSED METHOD

Web content can be regarded as an information source with hyperlinks and TV programs as another without them. To handle these different types of content, a method is necessary which can

measure both types of content in an integrated manner, not fully depending on conventional link analysis.

Figure 1 shows the processing steps of the indexing of the proposed method. First, topic keywords are extracted from the closed captions of TV programs, for example, by calculating the term frequency. Then, the candidate Web contents are obtained for each keyword using conventional search engines. By using conventional link-analysis-based search engines, the candidate Web content can be selected as information having a certain degree of quality. Common index is generated based on a vector space model using keywords from the Web content and the closed captions of TV programs. By referring to the common index, Web and TV programs can be ranked in an integrated manner based on a query given by the user. Figure 2 shows two types of search achieved by the proposed method. The keyword given by the user can be a query for integrated search to provide a mixed search result of Web and TV programs. Each search result can be a new query for chain search to provide related content. Chain search is done by computing similarity between the selected result and all other content based on the common indices.
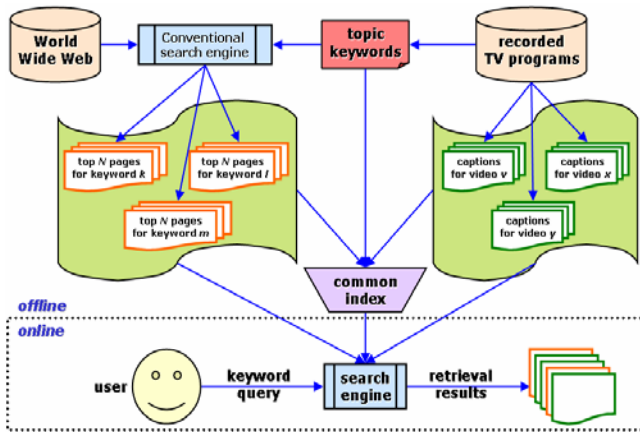

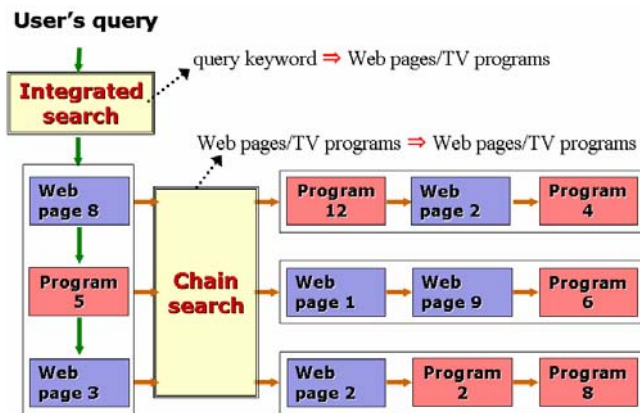**Figure 1. Processing steps of indexing**

Figure 2. Integrated search and chain search

3. IMPLEMENTATION OF PROTOTYPE

A prototype search engine based on the proposed method has been implemented. As shown in figure 3, an integrated search result of Web and TV programs are given as a list of text. When the user selects and zooms in the explanation text of a Web page or a TV program, the preview image of the page or the program

appears and its size can be changed smoothly. Also, the explanation text can be switched from one to another with a different level of detail according to the zooming. When the preview image becomes a certain size, chain search is triggered to provide content related to the focused content. Figure 4 shows an example of display transition when the user selects a TV program from a result of integrated search. The related contents obtained by the chain search are shown below the TV program. The zoom-based display is based on Zooming Cross-media[1], which provides an integrated framework to handle various content having different media and different levels of details. Testing of the prototype confirms that the proposed method achieves integrated search for Web and TV contents that match with the given query, and that users can efficiently acquire information according to the context by using chain search.
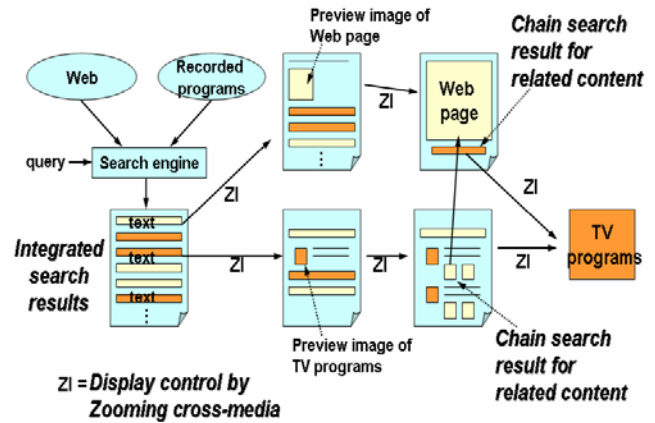

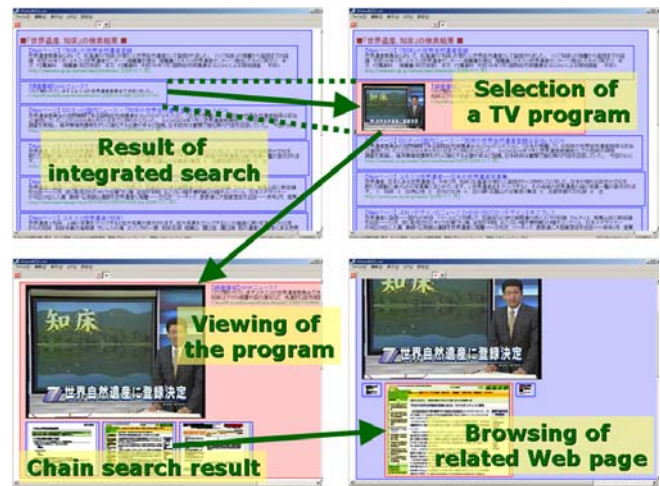**Figure 3. Browsing operation of search result by using Zooming Cross-media**

Figure 4. Display transition of implemented search engine

4. REFERENCES

[1] Araki, T., Miyamori, H., Minakuchi, M., Kato, A., Stejic, Z., Ogawa, Y., Tanaka, K.: Zooming Cross-Media: A Zooming Description Language Coding LOD Control and Media Transition. Proc. of the 16th International Conference on Database and Expert Systems Applications (DEXA2005), LNCS3588, pp.260-269, 2005.