A WWW-based System for Supporting Information Retrieval in Educational Uses

Itamiya Yuuki

School of Information Science,
Japan Advanced Institute of Science and Technology

February 14, 1997

Keywords: WWW, education, associative retrieval, information retrieval, information filtering.

The idea of using computers for educational environment was first proposed in the middle of 1950s. This idea was known as the area of CAI (Computer Assisted Instruction). The aim of CAI is to use the computer directly to support human's learning. Studies of CAI have been actively made until the present time. Recently with the appearance of high-performance and low-priced computers, a computer has become to be a common instrument which is widely used and there has been rapid development of the Internet. In the present situation, the computer have barely been used for CAI. Instead of CAI, the computer becomes a tool for communication and a tool for supporting thinking and expression. For example, a student uses a computer as a means for making presentation or getting informations scattered in all over the world. Especially, a well known application named WWW attracts a lot of interest of scholars and teachers in utilizing it for education purposes. There are two methods of using WWW for education. One is the method as done in CAI with networks. The other is the method of using WWW as a information resource of educational searches.

The education using WWW is expected to be popular in ordinary scene of education, in junior high schools or high schools, in the near future. In this thesis, I propose a support system based on WWW information retrieval for educational uses. The aim of the implementation of this system is to support 'learning by investigation'. In this paradigm, students learn by themselves through investigation of various media, a newspaper, a dictionary or an encyclopedia. Specially this system supports a student to get information from WWW. However, to realize this, there are some problems in existing systems based on WWW information retrieval (for example, various WWW search engines or support systems of information retrieval) as below.

- There are many WWW search engines existing in the Internet. This requires students to search each WWW page at a time and then several accesses are needed.

- Most of WWW search engines apply keyword matching method. But junior high school or high school students are not familiar to this method.

- There are several lists of searching results generated by WWW search engines. Students are reluctant to look at all of them.

- The result of searching by WWW search engines is not always reflected the present status of WWW pages. Therefore, useless information may be included in the result.

I propose solutions to these problems as follow. First, to support searching by keyword, I introduce the use of thesaurus of 'Word List by Semantic Principles' and associative words based on vocabulary of junior high school and high school textbooks of science curriculums and social curriculums. Those associative words are constructed using ' associative keyword vector' which was made by computation of cooccurrence probabilities of words and distances between words. Next, to support information retrieval using WWW, I propose a method to access several WWW search engines at the same time (The maximum is five engines) and display the result from these search engines selected by user together by merging and arranging. Next, I propose two information filtering processes to improve reliability of searching results. One is a process to investigate existence of WWW pages in result URL (Uniform Resource Locator) lists. The other is a process to investigate contents of WWW pages in the result URL lists. The former is a process which WWW robots investigate the WWW page and delete their data in the results, if the WWW pages are not in existence or can not access WWW server belong it. After that, the system presents the results of filtering. The latter is a process which WWW robots investigate WWW pages in the results in addition to the former process. This process analyzes the HTML (HyperText Markup Language) texts of WWW pages based on HTML tags. The system makes summaries and calculates scores of WWW pages based on the searching keyword. The final result is presented in the form of a list of short summaries ranked in the order of relevance scores.

An experimental system with the above functions is made. This system can be accessed from a WWW browser. The performance of the system is evaluated by a psychological method when subjects operate the system.

In the measurement of performance of the system, I measured the system's execution time. Those are times of each step of searching associative words, searching WWW pages through some WWW search engines and processes filtering the searching result. The aim of the measurement is focused on waiting times which users feel under practical uses. In the result of measurement, The average execution time of searching associative words is around five seconds. The average time of searching WWW pages is around a minute. The average time of process of filtering is almost two or three minutes. This result shows that the system is sufficient for supporting WWW information retrieval within practical time.

The experiment was carried out in two environments : with and without the proposed system. Subjects are told to search useful WWW pages for the investigation of a given theme given ten minutes. In this experiment, I keep data of URL lists made by subjects, log files of operation of the system and results of questionnaires carried out after the experiment.

In the results of the experiment, the number of URL collected in ten minutes with the system increased by 29% in comparison with the case of no system. The number of different URL collected by all subjects with the system increased by 37% in comparison with the case of no system, too. These results show that the system enables to realize efficient WWW information retrieval within the limited time. From the result of questionnaires, subjects shows that the system is useful for the WWW information retrieval.

The problems of this system to be solved in future are as follows. Due to the fact that WWW robots in the system influence and increase the network load, it is necessary to improve

the quality of WWW robots. An associative dictionary should be improved to deal with various words which users input. I plan to test this system with students in classes of junior high schools or high schools, although subjects of the experiment of this study is currently graduate students.

In this thesis, I propose a support system of the WWW information retrieval. The system is able to support students in the process of searching by keywords via WWW search engines. The system is useful for filtering and removing useless results of URL lists got by WWW search engines. These functions are shown to be useful for WWW information retrieval by experiments.