1. Introduction

The WWW provides an important avenue where different forms of information such as news, stock quotes, product updates as well as research results are made available. However, the dynamic nature of the Internet has pose real challenges to web surfers to keep track of latest information constantly and efficiently. Traditional search engines enable users to retrieve potentially relevant web information, but they are insufficient to track and monitor web pages based on user’s interests. Recently, intelligent information retrieval systems have emerged to help users to browse, find, filter as well as monitor web information. Most of these existing monitoring systems are providing electronic news scanning and clipping service.

A web information monitoring system, known as WebMon, has been developed in the Division of Information Studies, to track any web information over the Internet. Apart from providing basic monitoring functions, WebMon also supports portion monitoring of web pages. This allows the user to define sections, paragraphs, or even sentences in web pages to be monitored.

The technique used for portion monitoring is based on the Programming by Demonstration (PBD) concept. In WebMon, the user can specify the web page to be monitored, select the monitoring function and state the monitoring frequency to be used. When changes to these monitored pages are detected according to the specified criteria, the updated results are automatically extracted and stored into the user’s personal folders. They can then be displayed using a web browser.

2. Monitoring Functions

In order to support web information monitoring effectively, four monitoring functions have been identified: date monitoring (based on the last updated date), keywords monitoring, link monitoring, and portion monitoring. These functions are briefly described as follows:
• **Date Monitoring.** It enables a user to keep track on an update of web pages. Generally, an updated date is stored with each web page. This date is not necessarily visible to the user. When a modification is made, this date is automatically updated. Therefore, the system carries out date monitoring by checking and detecting changes in this date. This type of monitoring is rather general and is only useful for awareness purpose to inform the user that the page has been updated.

• **Keywords Monitoring.** It makes use of keywords specified by the users to keep track on specific changes that may occur in the monitored web page. For example, a user might want to be updated on any changes only on the keyword “notebook” within a computer company’s web page. The monitoring system will then record the number of occurrences of the selected keywords in the monitored web page. Subsequently, if changes to these keywords are detected on the web page, the user will be notified. This technique is very useful for monitoring specific areas of interest by the users.

• **Hyperlink Monitoring.** It allows users to monitor any changes in any of the hyperlinks of monitored web pages. For example, business competitors might want to keep track on the latest products from a company. In this instance, they can track the number of links in a product web page. Any change in the number of links in the web page will indicate a possible update on the product. As such, the users can be aware of the latest products that have been marketed by the company. In addition, the monitoring system can also detect whether the hyperlinks of the monitored web page are dead or moved by the provider.

• **Portion Monitoring.** It allows users to monitor a specific portion of a web page for updates. A copy-and-paste operation can be used to define the portion to be monitored. The monitoring system will then store the pattern information of the monitored page in its database. When the web page is updated, the system will extract the information from the specified portion area from the updated web page and compare it with the stored pattern information for change detection.

The portion monitoring function is very useful for monitoring specific information within a web page. As most web pages organize its information in a structured manner, some users are interested in only a certain part of a web page rather than the whole web page. In addition, when changes are detected, the users will only be informed on the changes of the specified portion, which eliminates the need for the users to search through the whole web page to find the changes. Portion monitoring can also be extended to monitor the whole web page. Thus, the system is very flexible to support different functions to monitor different types of web information.

3. **System Architecture**

Figure 1 shows the client-server based system architecture of the WebMon system. The Java programming language is used for the applets and servlets. The client subsystem supports three major services: User Management Service, Folder Management
Service and Monitoring Service. A user-interface is also designed to support user inputs on specifying monitored web pages, monitoring functions and the frequency to check the updates of a web page. The user needs to make use of a web browser such as Netscape’s Navigator or Microsoft’s Internet Explorer for displaying the interface and browsing through the web pages.

The server subsystem makes use of Java servlets to handle users’ requests. It tracks the web pages the users are interested in periodically and updates the monitored results whenever the monitored web pages have been changed. It consists of the Pattern Extraction Process and the Update Checking Process. The Pattern Extraction Process extracts the monitored web page (original web page) information and generates matching pattern with necessary information for subsequent checking according to the specified monitoring function. When the system is activated to check the updates, the Update Checking Process will retrieve the latest web page from the web. It then generates the matching pattern of the latest web page. Two sets of matching pattern from the original web page and the latest web page that are saved in intermediate files are used for update detection. The system also contains an internal database for storing the user profiles and necessary page information. A web server is developed for users to interact with the system through web browsers.

Figure 1. System architecture of WebMon.