Monitoring Web Information using PBD Technique

Bing Tan, Schubert Foo and Siu Cheung Hui

School of Computer Engineering, Nanyang Technological University,
Nanyang Avenue, Singapore

Abstract

Many web users would like to be kept informed on the latest development of their interested areas over the WWW. To discover web information, web users need to constantly monitor certain web sites and web pages for related information. However, the dynamic nature of the web has made such monitoring task complicated and time-consuming. As most web pages organize its information in a structured manner, users may be interested in only a certain part of a web page rather than the whole web page. This paper proposes a web monitoring system, WebMon, to help users to track a specific portion of a web page for updates. The technique for implementing portion monitoring is based on Programming by Demonstration (PBD).

1. Introduction

It is getting more often now that the WWW is used as a source for news, stock quotes, sports news, weather information, company information and many other types of information. In order to obtain updated web information, users often monitor certain parts of the targeted web pages constantly. Although web browsers such as Netscape’s Navigator and Microsoft’s Internet Explorer enable users to access various information sources over the web, users still need to spend much time in locating the desired information from the targeted web sites. They have to go through all the necessary web pages by repeatedly specifying different URLs (Uniform Resource Locators) [1] or by selecting them from bookmarks. In addition, they are also forced to browse through the whole web page in order to find the desired information. As such, the web monitoring task is a time-consuming process and thus poses a serious problem to users who need to monitor web information. Moreover, traditional search engines such as Yahoo! [2] and AltaVista [3] are not very effective for the purpose of searching and monitoring web information. This is mainly due to the fact that they are unable to keep up with the number of web pages and information created or modified from time to time.

This paper proposes a web monitoring system, WebMon, which can be used to track web information automatically on behalf of users. The system supports portion monitoring to allow users to monitor a specific portion of a web page for updates. Portion monitoring is
very useful for monitoring specific information within a web page. The technique for portion monitoring is based on Programming by Demonstration (PBD) [4]. PBD has traditionally used as a method to help users generate programs by demonstrating what they want with a normal user interface rather than developing codes in a programming environment. The PBD concept has been applied to different application systems. Each PBD system employs a different technique for learning (or inferring) and executing tasks. Here, the PBD is used to allow users to copy-and-paste the monitored portion into a user interface. The system then learns from the monitored portion to detect updates. This paper discusses the implementation of the WebMon system and its portion monitoring technique.

2. Web Information Monitoring

In recent years, a number of intelligent agents [5] have been emerged for intelligent information retrieval [6]. Many information retrieval agent systems [7] have been developed for different applications [8,9,10,11] such as web navigation systems, information finding systems, electronic commerce systems, and information filtering systems. Apart from these intelligent agent systems, a new form of Internet monitoring systems has recently emerged. Two types of information monitoring services are available. They are the electronic news clipping services [12] and web information monitoring services.

In general, electronic news clipping services such as EntryPoint [13], CyberAlert [14], WebClipping [15] allow users to specify news information such as headline news, stock quotes, sports results and weather news from the web. The system then searches, filters and extracts news articles from the trusted sources and news groups based on the specified topics and keywords. The extracted information is then delivered to the users via a personal web page given by the service provider. However, the news services provided are limited to the information sources supported by the service. They mainly extract news from trusted sources such as online newspapers and journals or from informal opinions and gossips of online discussion groups such as Usenet newsgroups, e-mailing lists, web bulletin boards and forums. Users have to define their interests within this list of limited information. In addition to supporting electronic news clipping services, some Internet monitoring systems such as eWatch [16], CyberScan [17] and NetMind [18] also support web information monitoring. They allow users to track and monitor web sites or web pages they are interested in. The users are notified of the updates through e-mails or a personalised web page.

Portion monitoring can be achieved through the use of the PBD concept as illustrated in some other monitoring systems such as NetMind and Internet Scrapbook. NetMind is a commercial product and no documentation is available on its implementation technique. Internet Scrapbook [19] supports monitoring of specified portions of web pages specified by users. The monitoring technique used in the Internet Scrapbook is based on a structural approach. It was based on the premise that article heading and positions in the web pages are rarely changed. However, Internet Scrapbook can only achieve a satisfactory successful rate based only on headings and positions of the articles. In this research, the proposed WebMon system will enhance the technique by offering an alternative approach using textual content and structural tags of the HTML documents for extraction and analysis instead of the heading of a portion. Thus, the matching process proposed by the WebMon system can improve and enhance the performance of the portion monitoring function.
3. The WebMon System

Figure 1 shows the system architecture of the WebMon system, which is based on a client-server architecture. Java programming language [20] is used for developing applets and servlets for the implementation of the WebMon system. The client subsystem supports three major services: User Management Service, Folder Management Service and Monitoring Service. The Monitoring Service provides the necessary user interface to support portion monitoring. Users can specify monitored web pages and the frequency to check the updates of a portion. The server subsystem makes use of Java servlets to handle users’ requests. It tracks the portions that the users are interested in periodically and updates the monitored results whenever the monitored portions have been changed. It mainly consists of the Pattern Extraction Process and the Update Checking Process.

![System Architecture of WebMon](image)

Figure 1. System Architecture of WebMon

In portion monitoring, the PBD technique is used based on a copy-and-paste operation. With this, the user can demonstrate which portion to monitor by copying the section of the web page and paste it into a selection area of the user interface. The monitoring system will learn the portion selected by the user. The pattern extraction process then extracts the portion information from the monitored web page (or original web page) and displays the portion in the interface. The extracted portion is verified by the user to ensure the correctness. If the portion is not correct, the user needs to select a portion again and copy-and-paste it to the interface again. Once the portion is correctly extracted, the pattern extraction process will generate matching pattern with necessary information for subsequent checking. 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 selected portion from the latest web page. The matching patterns of the portions from the original web page and the latest web page are then compared for update detection. The system also contains internal databases for storing user profiles, folders, monitored web page information and matching patterns. A web server is developed for users to interact with the system through web browsers.

Figure 2 shows the user interface for portion monitoring that uses a two-frame structure. The left-hand frame provides user management and the specification of monitoring criteria. In addition, it is also used for folder management that the user’s personal folder is displayed as a tree structure. On the other hand, the user can input the URL link in the “Enter url” field of the left-hand frame of the interface. The right-hand frame is used as a normal browser to display web pages to users. Once the URL is specified in the URL link area, the corresponding web page will then be retrieved and displayed. The right-hand frame is also used to display the monitoring results.

For portion monitoring, the user specifies a specific section of a web page to monitor. As shown in figure 2, the normal web browser in the right-hand frame allows users to browse through web pages and web sites. When the user wants to monitor something of a particular web page, he/she can just highlight the portion of the web page from the web browser, copy and paste it into the “Selection Area” in the left-hand frame. In this example, four rows of the table for fund prices in the CMG First State Investments web page are copied and pasted into the “Selection Area”. When any changes are detected subsequently in this monitored portion, the system will then record the updates in the specified document item that can then be presented to the user later through a web browser.
3.1 Pattern Extraction Process

In portion monitoring, the textual content and structural tags of the HTML documents are used for analysis. Figure 3 shows the pattern extraction process for portion monitoring. The displayable content of the portion (which has been copied by the user) is retrieved as the text pattern of the portion. Page Text Pattern Generator generates the text pattern (displayable content) for the web page using the HTML source document. The Portion Comparator then compares the text pattern of the portion with that of the web page to identify the location and properties of the portion.

Three layout structure formats have been commonly used for a portion. They are table, list and plain text. The plain text refers to displayable content, hyperlinks and images located outside the tables and lists in a web page. Table and list are structured layout formats, while plain text is unstructured format. Tables and lists are identified as blocks when the web page is parsed during the pattern extraction process. If the selected portion is located within a table or list of the web page, the portion is called a structured portion. Otherwise, if the portion contains plain text or plain text together with part of a table (or list), the portion is called a non-structured portion. In the example shown in figure 2, the portion is a structured portion as the selected section is located within a table of the web page.

![Figure 3. Pattern Extraction Process](image)

Once the location and properties of the selected portion have been identified, the Pattern Generator generates the matching pattern that the Update Checking Process will be needed subsequently when checking for updates. The matching pattern includes the page information and portion information as well as the displayable contents of the portion and the web page. The matching pattern is saved as a binary file in an internal database. Figure 4 shows the matching pattern of the selected portion shown in figure 2. Page information includes the title of the web page, the base web site address and the web page address. Portion information contains tag structure, portion mode, links, headings of a portion, block number and item number.
A tag structure is used to define the position of a particular tag within the HTML document. A single tag structure is presented as: tag-name pos1 pos2. Tag-name is the string value of the tag such as HTML, /HTML, HEAD, /HEAD, BODY, etc. Pos1 is the character position of the tag with respect to the HTML document from the first character and pos2 is the character position of the tag with respect to the displayable content of the document from the first character of the displayable content. For example, “BODY 206 39” is a tag structure to show the position of the tag <BODY>. Pattern Tag Start and Pattern Tag End are tag structures of the starting tag and the ending tag of a portion. Portion Tags are used to store all the tags of a portion in chronological order according to the HTML document. Portion Mode of a portion indicates whether the portion is structured or not. Mode “1” or “2” refers to structured or non-structured portion respectively. In addition, Links of a portion record all the hyperlinks and images of the portion.

Heading1, Heading2, Block Number and Item Number are only available for structured portions. The contents between two consecutive blocks are labeled as Heading1 for the second block. If the block is the first block of the web page, the contents before the block are taken as Heading1 of the portion. Heading2 of the portion is the first row of the table or the first item of the list. Block Number of the portion is the location of the block starting from the first block according to the HTML source of the web page. Item Number indicates the sequence number of the row in a table or list for the portion. In figure 4, the portion is located in the second table of the web page according to the HTML source (see figure 7). Thus, the block number of the portion is assigned as 2. Heading1 of the portion is null. This is because the introductory information of the web
page is stored in a table and is labeled as block number 1. No content is found between block 1 and block 2. Only hyperlinks are listed in the Links of the portion since there is no image available in this portion. In addition, the extraction date is also recorded.

3.2 Update Checking Process

The Update Checking Process is invoked whenever a user wants to perform an immediate update on the monitored web page or an automatic update based on user specified frequency. To perform the task, the latest web page of the monitored portion is first retrieved from the web. The matching patterns of the latest web page will be generated. The Update Checking Process is shown in figure 5.

The Matching Process aims to determine whether the portion has been changed. It compares the matching patterns of the portions from the original web page and the latest web page based on the displayable content, links and the tag information of the portion. Displayable content and links are used to detect changes in the content of the portion. Tag information is used to find out the structural changes of the portion. In addition, insertions, movements and deletions of the monitored portions can also be detected. This is done with the help of the displayable content and the tag information.

Figure 5. Update Checking Process
In the matching process, it first checks whether the portion is a structured portion. Heuristics is then employed to identify the correct location of the monitored portion in the latest web page. The matching pattern of the updated portion is then generated for comparison with the original matching pattern for update detection. If the portion is structured, Block Matching Algorithm will be used. Otherwise, Text Matching Algorithm will be used.

In the block matching algorithm, three levels of compliance checking using heading1, heading2 and block number are applied to locate the portion in the latest web page. Each level of compliance checking is based on the previous level’s result. Figure 6 shows the block matching algorithm. Figure 7 illustrates the block matching process using the selected portion in figure 2 as an example. As shown in the matching pattern of the portion, the portion that the user has selected is row 2, 3, 4 and 5 (#2, #3, #4 and #5) in table 2 (block 2). Heading1 for the portion is null as there is no content between block 1 and block 2. Heading2 is the first row of the table that is “Fund Prices as at 15 Feb 01”. For level 1 compliance checking, it failed because no Heading1 is available for checking. In level 2 compliance checking, Heading2 for the portion has been changed to “Fund Prices as at 19 Feb 01”. Therefore, the algorithm cannot locate the correct portion with Heading2, which means that the level 2 compliance checking failed. Then, level 3 compliance checking is used to locate the portion. Since the block number of the original
portion is block 2, the algorithm locates the same number of rows (#2, #3, #4 and #5) in block 2 of the latest web page as the latest information of the selected portion.

Figure 7. Three Levels of Compliance Checking

Figure 8 shows the text matching algorithm. The displayable content of the portion in the original web page and the displayable content of the latest web pages are used to locate the portion in the latest web page. When the algorithm checks for the updates, it first tries to locate a portion with 100% match with the original portion in the latest web page. If it is successful, it implies that the original portion remains the same in the latest web page. Otherwise, the matching algorithm will try to search for the portion that has the highest text matching percentage with the original portion. As long as there are only minor changes to the text, the algorithm will be able to identify the correct portion in the latest web page. However, if major changes are made to the text of the portion, the algorithm may not be able to locate the portion accurately. If the algorithm is unable to find any portions with more than 65% match with the original portion, the algorithm will return the message on “fail to locate the portion”.
If changes are detected, the Output Generator stores the new matching pattern of the latest web page inside the database to replace the original matching pattern. Email delivery service will be activated to notify the user about the changes. It also updates the changes into the user’s personal folder. In addition, the previous version of the selected portion is also provided for users to compare it with the current version of the portion. The type of changes detected is also given to indicate if any changes have been occurred. For structured portions, content change, insertion, deletion or movement can occur within a table or list. For unstructured portion, only content change is given. If no change has been detected, the original matching pattern will be retained in the internal database.

4. Performance Analysis

An experiment has been carried to measure the performance of the portion monitoring function in WebMon. As the speed of the system depends very much on the traffic of the Internet, we only focus on measuring the accuracy of the monitoring function. In the experiment, web pages from three domains, namely Education domain, Business domain and Entertainment domain, under the Yahoo! directories were selected for the experiment. 10 web pages that are most likely to be targets for monitoring were selected from each domain. A total of 30 web pages were used. These web pages were chosen also based on the condition that they could be downloaded for web page analysis. In addition, different monitoring tasks had been defined to test the monitoring function. The test web
pages were monitored using a weekly checking frequency for a period of one month. The accuracy of portion monitoring is 96.4%. There are many reasons why portion monitoring sometimes fails. This could be caused due to incorrect extraction of the selected portion. If the user selection is inappropriate, i.e. when the selected portion is either too small or spread over several blocks in a web page, the pattern extraction process may sometimes unable to generate the correct HTML source for the portion. The monitoring function works very well if the structure of the web page is stable. However, if the structure of the web page changes during the monitoring period, the update checking process may fail. Moreover, the portion monitoring may also fail if the web page has been removed from the web site or the URL address of the web page has been changed.

5. Conclusion

This paper proposes a web monitoring system, WebMon, to help users monitor web information from a specified portion of web pages. Portion monitoring is based on the programming by demonstration technique. The WebMon system is based on a client-server architecture using Java language. The client subsystem provides a user interface for users to specify the monitoring information. Folder management and user management services are also provided. The server subsystem supports the actual monitoring work. It consists of the Pattern Extraction Process and Update Checking Process for the generation of matching patterns from the monitored web pages and subsequent update checking. The performance of the portion monitoring function has also been discussed.

References


