Web Information Monitoring For Competitive Intelligence

Bing Tan, Schubert Foo and Siu Cheung Hui
School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore

Abstract

The WWW has become one of the most important media for sharing information. Web information provides another emerging and important avenue and source of competitive intelligence (CI) for companies. CI is critical for companies to stay competitive in the marketplace. Apart from business users, there are other types of CI users such as technical users, causal users, news awareness users and others who would like to be kept informed on the latest development of their interested areas over the WWW. To discover web information, CI 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. This paper proposes a web monitoring system, WebMon, to help users monitor specified web pages for latest changes and updates in information. Four monitoring functions including date monitoring, keywords monitoring, link monitoring and portion monitoring are supported by the system. The performance of these monitoring functions is also evaluated.

INTRODUCTION

Competitive Intelligence (CI) refers to the legal and ethical collection of information about competitors’ activities in the marketplace. To stay competitive in the business world, it is always necessary for a company to take proactive measures in terms of formulating competitive strategies way before related events happen. The company should be aware of the product and business information of its competitors. Such intelligence can be very helpful to make the company more competitive to market its current products, identify new products for development and for entering into new business areas. Therefore, one important means for CI is to keep continuously updated of competitors’ information.

With the rapid growth of the Internet, the World Wide Web (WWW) has become one of the most important media for sharing information resources. Digital documents, programs, images, movies and various other types of resources can be made available over the WWW. To exhibit and publicise their products and services over the WWW, many companies have already set up their web sites and pages which contain business information such as product
listings, new releases, general company information and technical white papers as well as recent company news on mergers and acquisitions. Unfortunately, such information is extremely useful for their competitors to track to know the latest products and services the companies are providing. This can then be used by their competitors to market their own products and services to be more competitive. Therefore, apart from traditional news sources, the WWW has much business information that can be mined for more business intelligence.

In order to obtain business intelligence from web information, users often monitor certain parts of the targeted web pages constantly in order to find new and updated information. For example, business users often need to monitor a company’s web page for its products or services that are usually contained in a specific location within the web page. Although web browsers such as Netscape’s Navigator (Netscape Communications Corporation, 2001) and Microsoft’s Internet Explorer (Microsoft Corporation, 2001) 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) (McKim, 1996) 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! (Yahoo! Inc., 2001a) and AltaVista (AltaVista Company, 2001) 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 describes a web monitoring system, WebMon, that can be used to track web information automatically on behalf of users for competitive intelligence. The system supports four monitoring functions including date monitoring, keywords monitoring, link monitoring and portion monitoring. In this paper, we first describe CI users and analysis on different types of web information monitoring services. Then, monitoring functions and system architecture of WebMom are described. Finally, the performance analysis of the monitoring functions is given.

WEB INFORMATION FOR COMPETITIVE INTELLIGENCE

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. Different types of web sites contain different contents. Online news web sites contain the latest news information. Stock market web sites contain information on stock prices. Sports web sites contain sports information and results. Companies web sites contain company information and, sometimes, the sales and promotion information are also included. This web information is good sources to monitor for competitive intelligence. For example, the sales and promotions information of a particular travel agent in its web site may trigger the competitors to cut down their price to compete for customers.
However, apart from business users, there are other types of users who are also desired for competitive intelligence. In general, four types of CI users can be identified. They are business users, technical users, causal users, news awareness users and others.

- Business users – they would like to monitor business-related information such as finance news, promotion campaign and new products release from competitors.

- Technical users – they would like to monitor new technologies and research results from engineering fields. Online technical journals and electronic publications can provide such information.

- Casual users – they would like to monitor information such as the release of new songs, movies, computer games, job openings and property information that are of their general interest.

- News awareness users and others – they would like to be informed about what is happening in the countries and the regions in a timely manner. Some would also like to monitor the stock prices, sports results and weather information.

**INTERNET MONITORING SYSTEMS**

In recent years, a number of intelligent agents (Klusch, 1999) have been emerged for intelligent information retrieval (Yang & Pedersen, 1999). Intelligent information retrieval has become increasingly important to find the most relevant information to meet users’ need. Traditional search engines enable users to retrieve potentially relevant web pages, but unable to focus on the information need to a particular user group. In contrast, intelligent agents provide the personalised functions that automatically learn user interests, detect significant events, present the main content of the events, alert the users about the changes and monitor events of interests to the users’ need. They are developed to act on behalf of the users to support the activities automatically.

Different information retrieval agent systems (Mladenic, 1999) have been developed. Web navigation systems (Armstrong *et al.*, 1995; Lieberman, 1997; Shavlik *et al.*, 1999) “learn” user’s interest and guide the user browsing through the web based on it. Information filtering systems (Lang, 1995; Konstan *et al.*, 1997) filter top news items based on the user’s interest. Information finding systems (Kruilwich & Burkey, 1996; LaMacchia, 1997) search for relevant information from the web based on user profiles. Electronic commerce systems (BargainFinder Press Ltd., 2001; At Home Corporation, 2001) support different stages of commerce process including searching, negotiation and deal settlement. Some other agent systems such as CiteSeer (Bollacker *et al.*, 1998; Giles *et al.*, 1998) is designed for scientific publication retrieval on the web while Calendar Apprentice (Mitchell, 1994) is used for scheduling meetings.

Apart from the above agent systems, a new form of Internet monitoring systems have recently emerged. Two types of information monitoring services are available. They are
electronic news clipping services (ENC) and web information monitoring services (WIM). Table 1 provides examples of Internet monitoring systems. The service provided, monitoring functions, delivery mechanism and domain of these systems are listed in the Table. Internet monitoring systems such as My Yahoo!, Infogate, CyberAlert and WebClipping provide electronic news clipping services. Some other systems such as eWatch, CyberScan and NetMind provide both electronic news clipping and web information monitoring services.

### Electronic News Clipping Services

My Yahoo! (Yahoo! Inc., 2001b) is considered as a personalized version of Yahoo!. It clips information by topics and keywords that users have specified. However, users are limited to the choice of topics and news provided by the sources available by the system. The users can set up their personalized pages that can associate with different topics they are interested in. For Infogate (Infogate Inc., 2001), it acts more like a news broadcasting service. The users need to download and install the proprietary client software before using the system. News information supported includes headline news, stock quotes, weather news, soccer results, etc. Using the push-like technology, the system can proactively “push” or “broadcast” the news information in real-time to the users through a friendly user interface. Users can select the news channels, the type of information from each channel and the update frequency.

**Table 1. Internet monitoring systems**

<table>
<thead>
<tr>
<th>SYSTEM</th>
<th>SERVICE</th>
<th>MONITORING FUNCTIONS</th>
<th>DELIVERY MECHANISM</th>
<th>DOMAIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>My Yahoo!</td>
<td>ENC</td>
<td>Page Monitoring</td>
<td>E-mail and web page</td>
<td>Electronic news, TV Listings</td>
</tr>
<tr>
<td>Infogate</td>
<td>ENC</td>
<td>Page Monitoring</td>
<td>Web page</td>
<td>Electronic news</td>
</tr>
<tr>
<td>CyberAlert</td>
<td>ENC</td>
<td>Page Monitoring</td>
<td>E-mail and web page</td>
<td>Electronic news</td>
</tr>
<tr>
<td>WebClipping</td>
<td>ENC</td>
<td>Page Monitoring</td>
<td>E-mail and web page</td>
<td>Electronic news</td>
</tr>
<tr>
<td>eWatch</td>
<td>ENC, WIM</td>
<td>Keywords Monitoring and Page Monitoring</td>
<td>Fax, e-mail and web page</td>
<td>Online forums, Usenet news and bulletin board messages</td>
</tr>
<tr>
<td>CyberScan</td>
<td>ENC, WIM</td>
<td>Page Monitoring</td>
<td>E-mail and web page</td>
<td>Online publications, newspapers and magazines</td>
</tr>
<tr>
<td>NetMind</td>
<td>ENC, WIM</td>
<td>Page Monitoring, Text Monitoring, Precision Monitoring, Image Monitoring, Link Monitoring and Keywords Monitoring</td>
<td>E-mail and web page</td>
<td>Web sites and electronic news</td>
</tr>
</tbody>
</table>
CyberAlert (CyberAlert Inc., 2001) is an Internet monitoring and clipping system for public relation and market intelligence purposes. A wide range of on-line information sources including 425 news sites, normal web sites, web-based discussion groups and Usenet news groups are monitored in order to support the services. A personalized filing system of folders for different topics are given to each user or user group. The user can specify search terms in the form of keywords or phrases with Boolean operators. CyberAlert provides a good user interface, in which information can be sorted by many criteria such as search term, date or source. WebClipping (AllResearch Inc., 2001) is a clipping service that helps the users monitor on-line company-specific information such as company’s activities, products or services. It scans 650 online news sources including news archives, news wires, news indexes and 50,000 Usenet discussion groups for user specified information. When the user signs up the service, he/she can then enter the company profile and search terms. The system will perform a comprehensive run to gather information relevant to the company based on the search terms over the Internet. Once this is done, it will only clip newly updated information.

In general, electronic news clipping services (Johnson, 1999) such as My Yahoo!, Infogate, CyberAlert, WebClipping 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.

**Web Information Monitoring Services**

eWatch (PR Newswire, 2001) is an on-line monitoring service of public discussion. Focusing on Usenet groups and other e-lists only, eWatch is normally used to find what consumers and professional are saying and reading about a specific company's services, products and competitors. Besides, eWatch supports a web site tracking agent called WebWatch. The agent automatically checks the monitored sites (specified by the user) once a day. The user can provide filtering keywords to specify the interested information. CyberScan (CyberScan Technologies, 2001) provides three sub-services: WebScanning for monitoring web pages, Opinion Monitoring for monitoring Usenet newsgroups, electronic mailing lists, web bulletin boards as well as forums and Site Tracking for monitoring web sites. For the first two services, the user can specify keywords, topics and other search parameters. Site Tracking service allows users to specify up to 10 web sites to monitor for relevant changes. The system notifies the users the results by e-mail or provides a password-protected web-based report.

NetMind (NetMind Technologies Inc., 2001) is a commercial product which monitors web pages that the user selects and notifies the user by e-mail when the specified page has been updated and changed according to user specification. NetMind has six major components
including Page Minding, Text Minding, Precision Tracking, Image Minding, Link Minding and Keywords Minding to track changes on web page, text, image, link and keyword of web pages. However, there are two problems with the system. First, the system response is very slow, this is mainly due to its increasing number of users. Second, the system sends e-mails to users with a link to the updated page without highlighting the changes on the page. The user still needs to browse through the whole web page in order to find out the changes that have been made. This is sometimes quite difficult and time-consuming.

Thus, in addition to supporting electronic news clipping services, some Internet monitoring systems such as eWatch, CyberScan and NetMind 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. However, these systems only provide some primitive support for monitoring changes in web sites and web pages. Generally, basic monitoring functions of these systems are quite limited. As they are commercial systems, no documentation is available to allow a more thorough discussion of these systems.

**THE WEBMON MONITORING SYSTEM**

In this research, the WebMon monitoring system is proposed for tracking any web information over the Internet. Apart from providing basic monitoring functions, WebMon also supports portion monitoring of web pages. The technique for portion monitoring is based on the Programming by Demonstration (PBD) (Cypher, 1993) 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 the monitored web pages have been updated or changed according to the specified criteria, the updated results are automatically stored into the user’s personal folders and documents. They can then be displayed using a web browser. At the same time, email notifications are sent out to inform the users about the updates.

**WebMon Monitoring Functions**

In order to support web information monitoring effectively, four monitoring functions have been identified: date monitoring (for last modified date), keywords monitoring, link monitoring and portion monitoring. These are described as follows.

- **Date Monitoring**. This function is to monitor the last modified date of the web page. 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 occurred 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 and positions of the selected keywords in the monitored web page. Subsequently, if any changes have been detected on the web page, the user will be notified. The changes include the original keywords that are deleted or new keywords that have been added. This technique is very useful for monitoring specific areas of interest by the users.

- **Link 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. They can track the number of links occurred in the 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 information 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. With this, the user can copy a section of a web page and paste it into a selection area of the user interface. 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.

**System Architecture**

Figure 1 shows the system architecture of the WebMon system which is based on a client-server architecture. Java programming language (Arnold & Gosling, 1998) is used for 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. It is developed as Java-based applet programs. 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 that 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. The matching patterns 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.

All the services and processes of the system work with each other to fulfill the monitoring task. However, each of them has its own responsibility and functions. Here, we will describe these services and processes in detail.

**User Management Service**

User Management Service maintains user information and provides the necessary functions needed by the user to interact with the system on updating his personal information. The
functions include creating a new user, registration of the current users, verifying user information and updating the personal information in the user profile. A database is created to store users’ information. Whenever a user logs into the system, the User Management Service checks the user’s status. A current user just needs to inform the system by providing the user ID and password. The inputs are then verified. If a new user logs into the system, user particulars such as user ID, full name, e-mail address and other necessary information are required to create a new account for the new user. The system checks all the inputs and makes sure that they are in correct formats. If there are no problems with the inputs, a new record for the user will be created in the user profile. The user cannot use the system until all the information is verified. A notification will be sent to the user to confirm the registration. In addition, user particulars can also be updated through the user interface if necessary.

Folder Management Service

The Folder Management Service is used to manage the personal folders of users. The WebMon system supports multi-user access. Each user is allowed to add, delete and move folders and documents that contain different monitored items within his/her personal folders. The database residing at the server subsystem will be updated accordingly when a modification is made. The folders are organised as a tree-like structure and can be navigated through the user interface. The tree structure of a user’s personal folder is shown in figure 2. The representation of the folder hierarchy is in the form of node types. The node type can be

![Folder Structure Image]

Figure 2. Tree structure of a user’s personal folder
classified as root, folder, document and item. The root is a single node containing the identity of the user. The sub-nodes will be folders with multiple documents or sub-folders inside. Folder nodes can contain sub-folder nodes or documents. Document nodes contain item nodes. The item nodes contain individual monitored web information that the user wants to track. Users can create folders and documents to organize their monitored information in the personal folders.

Monitoring Service

The Monitoring Service provides the necessary user interface to support the four monitoring functions including date monitoring, keywords monitoring, link monitoring and portion monitoring. Figure 3 shows the user interface for portion monitoring. In this case, a two-frame structure is used. The left-hand frame provides user management, all the necessary monitoring functions 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 monitored results. For date monitoring, keywords monitoring and link monitoring, the users just need to
indicate the type of monitoring they want, the monitoring frequency and the corresponding input such as the keywords and the URLs for hyperlinks.

For portion monitoring, the user can specify a specific section of a web page to monitor. As shown in figure 3, 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. With the programming by demonstration concept, it allows the user to copy-and-paste the monitored portion into the user interface. The system then learns from the monitored portion to detect updates. As discussed in the Folder Management Service, each document can contain multiple monitored items. Each portion is regarded as an item to be monitored by the monitoring system.

**Pattern Extraction Process**

When the user selects a web page from the monitoring service, the Pattern Extraction Process parses the web page according to the monitoring functions. If date monitoring, keywords monitoring or link monitoring is invoked, the pattern extraction process will extract the necessary information about the last modified date, keywords and hyperlinks from the web page for subsequent checking for updates. The URL address of the web page is also saved. For example, in keyword monitoring, the locations of the specified keywords and occurrence of the keywords are detected and stored as meta-information.

However, it is more complicated for portion monitoring which adopts the copy-and-paste technique. The textual content and structural tags of the HTML document are analysed in
order to extract the correct and complete information from various parts of the whole HTML document to ensure the integrity of the portion been monitored. This is necessary since the required information may be scattered in various sections before the start of the defined portion.

Figure 4 shows the pattern extraction process for portion monitoring. The displayable content of the portion 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 3, the portion is a structured portion as the selected section is located within a table of the web page.

Once the location and properties of the selected portion have been identified, the Pattern Generator generates the matching pattern. This is subsequently used by the Update Checking Process to check 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 5 shows the matching pattern of the selected portion shown in figure 3. The page information includes the title of the web page, the base web site address and the web page address. Portion information contains the tag structure, portion mode, links, headings of a portion, block number and item number.
Figure 5. An example matching pattern

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 5, the portion is located in the second table of the web page according to the HTML source (see figure 9). 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.

Figure 6 show the contents of the portions after the extraction process. In this example, two items are stored in the document “companies”. The “cmg” item refers to the portion that has been selected in figure 3. The “cNet” item contains the portion that had been extracted from the home page of CNET Networks. The item names, “cmg” and “cNet”, are specified by the user. In addition, the title of the web page, the link to the web page and the date of extraction are also displayed together with the portions.
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 Update Checking Process is shown in figure 7. For date monitoring, keywords monitoring or link monitoring, the necessary information such as the last modified date, keywords or hyperlinks of the latest web page is extracted. This is then compared against the original information to detect changes. For portion monitoring, the matching patterns of the latest web page will first need to be generated.

The Matching Process aims to determine whether the web page content 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.

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
The matching pattern of the updated portion is then generated for comparison with the original matching pattern for update detection. If the portion is a structured portion, the block matching algorithm is used, otherwise, the text matching algorithm is used.

![Flowchart](image)

Figure 8. Block matching algorithm.
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 9 shows the block matching algorithm. Figure 9 illustrates the block matching process using the portion item “cmg” 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 reads “Fund Prices as at 19 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 implying that the level 2 compliance checking has also failed. Level 3 compliance checking is then 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 10. Text matching algorithm

Figure 10 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 a 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 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 status of “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. An 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. Figure 11 shows the interface that displays the two versions of the monitored portion for the item “cmg”. The type of changes detected is also reported by the system. For structured portions, content change, insertion, deletion or movement can occur within a table or list. For unstructured portions, only content change is given. If no change has been detected, the original matching pattern will be retained in the internal database.

**PERFORMANCE ANALYSIS**

An experiment has been carried out to measure the performance of the monitoring functions 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 functions. The system was first tested using a set of created web pages by introducing various changes to confirm the correctness of the system functionality. A subsequent real life test was carried out. In this respect, an earlier study was carried out by the same authors to analyse the characteristics of
web page updates over 7 domains on the web (Tan, Foo & Hui, 2001). This was used to identify different characteristics of 105 web pages over a one-month period, including total and average content change by domains, average content change by web page types, by web page attributes and page change frequencies. In the experiment, the 3 most active domains from this study, namely, the Education, Business and Entertainment domain, under the Yahoo! directories were chosen. A total of 10 web pages from each domain, making up a total of 30 web pages, were monitored by the system using a weekly checking frequency for a period of one month. Thus, a total of 120 checks were made by the system. Table 2 shows the results of the performance analysis.

Table 2. Performance results for the monitoring functions

<table>
<thead>
<tr>
<th>Function</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portion Monitoring</td>
<td>Pattern Extraction Process 93.3%</td>
</tr>
<tr>
<td></td>
<td>Update Checking Process 96.4%</td>
</tr>
<tr>
<td>Keywords Monitoring</td>
<td>100%</td>
</tr>
<tr>
<td>Link Monitoring</td>
<td>100%</td>
</tr>
<tr>
<td>Date Monitoring</td>
<td>100%</td>
</tr>
</tbody>
</table>

For date monitoring, keywords monitoring and link monitoring, the accuracy reaches 100\% implying that these monitoring functions work perfectly in the experiment. For example, in date monitoring, changes to the date of the web pages were correctly detected in all 67 occasions when it occurred. The same performance was achieved for both keywords monitoring and link monitoring functions.

For portion monitoring, the performance was measured based on the two major processes: pattern extraction process and update checking process. During the pattern extraction process, 2 out of 30 web pages failed in the extraction process, thereby yielding an accuracy of 93.3\%. The reason of incorrect extraction of the selected portion is mainly due to “inappropriate” user selection in which the pattern extraction process is unable to generate the correct HTML source for the portion. In these two failed cases, the selected portions are from a part of a table row (or cell). However, the pattern extraction process at present can only process a complete row of table, which is the smallest unit of a table treated by the process.

The other 28 web pages were then used for performance measurement in the update checking process. The update checking process had been invoked 112 times during the month. The process failed in 4 cases yielding an update checking process accuracy of 96.4\%. The monitoring function works perfectly if the structure of the web page is stable. However, if the structure of the web page changes substantially during the monitoring period, the update checking process may fail. This is because the portion of the new web page that corresponds to the old one may not be located correctly. For the 4 failed cases, major changes were encountered in the structure of the monitored web pages. The portion monitoring will obviously “fail” if the web page has been removed from the web site or if the URL address of
the web page has been changed. In this case, the process will inform the user with the message “fail to load the web page”. These problems are currently under further investigation to facilitate a solution for them. Nonetheless, the accuracy for portion monitoring is credible and should still prove useful in almost all cases of portion monitoring even at its present state.

CONCLUSION

The WWW contains much useful information that can be discovered by companies for business intelligence to stay competitive in the marketplace. To achieve this, this paper proposes a web monitoring system, WebMon, to help users monitor web information. The WebMon system is based on a client-server architecture using the Java language. The client subsystem provides a user interface for users to specify the monitoring information. The monitoring functions supported include date monitoring, keywords monitoring, link monitoring and portion monitoring. 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 monitoring functions has been analysed. As part of the evaluation work, we are now conducting a user evaluation on utilities test of the system in order to provide more efficient and effective monitoring functions for competitive intelligence.

REFERENCES


McKim, G. W. 1996. Internet research companion. Indianapolis, IN: Que E&T.


