Web Information Monitoring: An Analysis Of Web Page Updates

Bing Tan, Schubert Foo and Siu Cheung Hui

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

Abstract

The rapid growth of the Internet has been so astounding that web information can be changed and updated at any time. This dynamic nature of information content on the web has posed a serious problem to users who need to constantly keep track of the latest updates on some specific information. Traditional search engines enable users to retrieve potentially relevant web information, but they are ineffective to track and monitor web pages based on users’ interests. On the other hand, web information monitoring systems are designed specifically to help users track and monitor web information. However, to make web monitoring effective, it is necessary to identify and understand typical web page update characteristics so that useful monitoring features and functions can be designed and built into these systems. In this study, a total of 105 web pages from the Internet are collected and monitored over a one-month period. These pages are selected from seven domains that include Art, Education, Business, Entertainment, Science, Social & Culture, and Government under Yahoo!’s directories. The analysis results are presented according to web site domains, web page types, web page attributes and change frequency. Based on this study, different functions and features for a web monitoring system are identified. These features have been incorporated into a web monitoring system, WebMon, that has been developed at School of Computer Engineering, Nanyang Technological University, Singapore.

Introduction

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 accessed through the WWW. Research results, new findings, and even opinions can also be put on the web from anywhere at anytime. Due to the dynamic nature of web information, the contents of many web sites and its web pages are constantly being updated to reflect the latest changes.

Generally, when users surf the WWW, they are either searching for specific information or simply browsing the web for something interesting. It is quite usual now
for users to use the WWW as a source for news, stock quotes, weather information, product updates and other business-related information that are being made available daily. For example, business users would like to monitor business-related information such as finance news, promotion campaigns and new products release from competitors. Such information is essential for them to maintain the competitive edge of their companies. Technical users would like to monitor new technologies and new research results from engineering fields. Online technical journals and electronic publications can provide such need. Casual users 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 would like to be informed about what is happening in the countries and the regions in a timely manner. Yet others would like to monitor stock prices, sports results and weather information. In other words, users are interested in a variety of information from different sources and there is a real need for systems to be developed to support the task of automatically identifying changes in the information and making them known to users in a timely and effective manner.

Although web browsers such as Netscape’s Navigator (Netscape Communications Corporation, 2000) and Microsoft’s Internet Explorer (Microsoft Corporation, 2000) 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 monitoring of web information is a time-consuming process and thus poses a serious problem to users who need to constantly access some web pages for the latest updates of relevant information.

Traditional search engines such as Yahoo! (Yahoo! Inc., 2000a) and AltaVista (AltaVista Company, 2000) are not very effective in 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 every day. A number of monitoring services have been developed to monitor on-line news information. The most notable electronic news monitoring services include My Yahoo! (Yahoo! Inc., 2000b), CyberAlert (CyberAlert Inc., 2000), WebClipping (AllResearch Inc., 2000) and Entrypoint (EntryPoint Inc., 2000). These systems 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. In addition to news monitoring services, some web monitoring systems such as NetMind (NetMind Technologies Inc, 2000), CyberScan (CyberScan Technologies, 2000) and eWatch (PR Newswire, 2000) are also available to provide some primitive support for monitoring web information. However, the supported monitoring functions are insufficient for most users.
In this paper, we focus on investigating the update characteristics of web page contents. This work forms part of our research in the development of a web-based monitoring system, WebMon (Tan et al., 2000), for monitoring and tracking web information automatically. Thus, the motivation for this study is to identify the various monitoring functions, features and characteristics of web page information for the design of WebMon. Such information is lacking and is especially useful since it is also applicable to the design and research into other forms of web information monitoring services. The rest of this paper is organized as follows. The definitions of some important terms that help to understand the updating patterns are first provided. The web page analysis procedure is described and the results of the analysis are then presented. The monitoring features that have been identified from the study are discussed and the status of WebMon development is reported.

Web Page Update Analysis Procedure

In this section, a number of important terms are first discussed to set the context of the web page update analysis procedure and study findings. These terms include web site domains, web page types and web page attributes.

Web Site Domains

Domain indicates the “pedigree” of a web site. There are several ways to categorize domains of web sites based on the URLs (Koehler, 1999). The URL consists of two or more URL fragments separated by dots. The top-level domain, the right-most tag, may indicate a functional or geographic domain as follows:

- Functional domain. It indicates the type of the publisher for the web site. For example, “.com” indicates commercial, “.edu” for educational, “.gov” for government, “.org” for non-government organisation and “.net” for network provider.

- Geographic domain. It uses the last two-letters of the URL address to identify the country and region of publication. Sometimes, this is used in conjunction with functional definitions to indicate the type and location of the publisher. For example, the fragment “ac.uk” indicates a British academic server, and “co.jp” signifies a commercial Japanese web site.

Yahoo! has further extended the functional definitions to a total of 14 directories in its portal page to define different domains. In this study, only web pages from seven directories, which include Art, Education, Business, Entertainment, Science, Social & Culture and Government, are selected for investigation.

Web Page Types

Each functional web site usually comprises a number of web pages linked by hyperlinks. Hass and Grams (Haas and Grams, 1998) gave a general classification based on the
The functional purpose of the web pages. The classification contains 7 major types as follows:

- **Organisational Page**: This type of web pages normally provides some kind of index structure or access aids to the users. This includes publication index pages that contain the table of contents of all the documents, web site content pages that display links to available pages or sub-pages within a single site and index pages that provide links to documents, other sites, pages and sections of a web site.

- **Documentation**: This type of web pages is mainly used for reference purposes, in the sense that users are likely to look up a certain part of the web page to get information on some specified queries. The most common form of this type of web pages is FAQ pages. Others include web pages containing descriptions of services or products.

- **Text**: This refers to a very broad class that includes most types of writing. Examples include articles from newspapers and magazines, scholarly papers, contracts, bibliographies, resumes, etc.

- **Home Page**: This refers mainly to the home page of a company, organisation or person. This type of web pages may have certain regularity in structure. Information contained in these web pages include introduction to an organisation or person, organisational structures, brief news articles and descriptions of products or services.

- **Multimedia**: These web pages include sound, video, image, graphics and other non-textual documents. For example, a user can view a video clip of online news on the web page.

- **Tools**: These web pages provide the means for the users to accomplish a task online. For example, a user can use a search tool for seeking information. Order forms, e-mails or comment forms are other examples of tools.

- **Database Item**: This refers to web pages that contain highly structured information, as might be found in a database. For example, an online bookstore web site contains web pages displaying titles, authors and price information of books in a structured manner.

All these defined page types, or more likely, combinations of page types, can be found in individual web pages. As such, a page might be identified as containing one or more primitive types rather than falling into a single type. In addition, the classification is not claimed to be exhaustive because it is done only from the sample web pages collected. Given the rapid growth of the web, newly created web pages may include other web page types that are not covered in the above taxonomy.
**Web Page Attributes**

Web page attributes (Chan and Yu, 1999) include elements within a page such as text, hyperlink, image, color, etc. Text refers to words and paragraphs of textual contents within a web page. Hyperlink represents a relationship between two web pages or just between sections of the same page. When clicking on a hyperlink, the user will be led to another web page or section of the page. Image is a graphical file to be displayed. Some of the image design elements include banners, horizontal lines, bullets, etc.

**Hypertext Markup Language**

Most of the documents on the WWW are written in HyperText Markup Language (HTML) (Ladd and Donnell, 1996). HTML is a standard language for describing the contents and appearance of web page on the WWW. The WWW consortium or W3C (World Wide Web Consortium, 2000) was established to maintain standards for HTML and related technologies. The latest version is HTML 3.2 and the W3C is currently recommending HTML 4.0 based on the earlier work on HTML tags.

**Analysis Procedure**

A total of 105 web pages were collected to study the web page and web site behavior over a one-month period. This is aimed to collect data as a means to study the frequency of information changes of these web pages, the types and characteristics of changes in web page contents. Findings from these samples collected will enable a better understanding of desired features and functions of a monitoring service.

The sample web pages were selected from seven domains under the Yahoo!’s directories, namely, Art, Education, Business, Entertainment, Science, Social & Culture and Government. An average of 15 web pages was chosen from each domain. In addition, care was taken in selecting web pages so that they are representative of the information found in their respective domain areas.

In this study, Microsoft’s Internet Explorer 5 was used to download the web pages. To ensure data consistency, once data collection had started, no new web pages were added into data collection. In order to monitor the behavior of web pages, the changes of the selected web pages were recorded on a daily basis. The Symantec’s Visual Café (Jardin and Dixon, 1997) text comparison tool was used to parse the HTML source of the web pages for change detection. In this study, the collection of web pages was monitored for changes based on web site domains, web page types, page attributes and update frequency. The analysis was performed based on the analysis of two corresponding sets of HTML codes from each source web page. Changes were identified based on the following analysis of the HTML codes:

- Text Analysis. This refers to the analysis of changes on textual contents that could be displayed in a web page. In text analysis, the tag sequence for each web page is examined. The tags are represented as `<p>`, `</p>`, `<font>`, `</font>`, `<br>`, `</br>`, `<tr>`, `<t>`, etc.
• Link Analysis. This refers to the analysis of changes in links of the web page. In link analysis, the tag sequence of the web page is examined for link tags. For example, the tag sequence of `<A href= "http://www.ntu.edu.sg/main.html">Nanyang Technological University</A>` is used to represent a hyperlink and the linked target destination. Here, “Nanyang Technological University” is the displayable text of a link on the web page. When clicking on “Nanyang Technological University”, the user will be led to the web page with URL of `http://www.ntu.edu.sg/main.html`. Changes within this link content will be treated as an update.

• Image Analysis. This refers to the analysis of the image content of the web page. In image analysis, the tag `<img>` is used to identify the name and type of the image file. For example, `<img SRC="notecards.jpg" BORDER=0 height=117 width=150>` indicates that the name of the file is “notecards”, the type of the file is “.jpg” and the height and width parameters define the size of the image. Any subsequent changes in the file name, type and size of the image will be treated as an update.

• Layout Analysis. This refers to the analysis of the layout of the web page in terms of the structure of the web page, background color, background image and fonts. Tags such as `<bgcolor>`, `<background>`, etc. provide the necessary information about the layout change of a web page.

• Last Updated Date. This refers to the analysis of the web page update that is usually reflected in the last updated date field of the web page. The last updated date informs the users on how recently the information on a web page has been changed. For example, “Last updated on 29/3/2000” may be displayed anywhere of a web page. However, last updated date usually appears at the top or the bottom of a web page.

**Analysis of Results**

After the 105 web pages were collected and monitored for a one-month period, they were analysed and results were obtained. In this section, we first give an overall observation of the web pages that were collected. Two new terms are defined:

• Average Content Change. This indicates the average changes on content of a web page for each change. In this definition, changes in web page attributes include those of text, hyperlink, image, and last updated date. For example, if there are 16.3% of text change, 8.4% of hyperlink change and 3.1% of image change for a web page on a particular day, the content change percentage will be 27.8% (16.3% + 8.4% + 3.1%) for that web page on that day. The percentage of text change effectively distinguishes the actual number of word changes between two versions of a web page. A 16.3% text change implies that 16.3% of the total number of displayable words in the original web page has been changed in the new web page. In the study,
the average content change will be the average of content change for all the web pages for the whole monitoring period.

- **Total Changes per Page.** Web page change indicates a binary result on whether a change of an attribute has occurred on a web page. Any change, regardless of how much, for an attribute that has occurred on a web page is considered as one change for that attribute. For example, if text change, hyperlink change, last updated date change and image change have occurred for a web page on a particular day, the total changes for the web page are counted as 4 on that day. The total changes are the sum of all the changes for one month’s monitoring period. Total changes per page are calculated as the average of the total changes of all the web pages divided by the total number of web pages over the monitoring period.

Generally, the number of total changes per page is a more useful parameter than average content change of a web page for monitoring purpose. It describes the web page update behavior over time rather than measuring how much of the content has been changed in a web page.

**General Observation**

Table I shows the general features of the collected web pages. The presented data is mainly based on the web page attributes such as hyperlink, image and last updated date. As a result, from the 105 collected web pages, 96.1% of the web pages contains hyperlinks, 78.6% contains images and 43.6% contains last updated date. Apart from textual contents, hyperlink and image are other common attributes that most web pages have. It is also observed that images sometimes contain hyperlinks so that when a user clicks on the image, it will link to another web page. In addition, the last updated date tends to appear at the root pages or branch pages rather than at the leaf pages.

<table>
<thead>
<tr>
<th>FEATURES</th>
<th>PERCENTAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pages with Hyperlink</td>
<td>96.1%</td>
</tr>
<tr>
<td>Pages with Last Updated Date</td>
<td>43.6%</td>
</tr>
<tr>
<td>Pages with Image</td>
<td>78.6%</td>
</tr>
<tr>
<td>Dynamic Pages</td>
<td>3.8%</td>
</tr>
<tr>
<td>Use of Java Applets</td>
<td>12.6%</td>
</tr>
<tr>
<td>Use of JavaScript</td>
<td>7.7%</td>
</tr>
</tbody>
</table>

The majority of web pages are authored using the tagged standardised HTML language. In addition, Dynamic HTML techniques can also be used for web page authoring thereby making it more dynamic and interactive. However, there is no standard for Dynamic HTML. Different forms of Dynamic HTML exist. For example, Dynamic HTML may include style sheets or cascading style sheets that are incorporated in the Web browser object model so that Javascript & VBScript can be used to alter these styles.
dynamically. Downloadable fonts can be used and changed dynamically. Likewise, the web page contents can also be changed dynamically. Nonetheless, the percentage of web pages that uses Dynamic HTML is still quite low. In this study, only 3.8% of the web pages of the sample collected are Dynamic HTML pages.

Currently, Java (Arnold and Gosling, 1998) is one of the most commonly used programming languages for web applications. It allows applets and servlets to be run on the client side and server side respectively. Web pages are increasingly being developed with Java applet and JavaScript (Lemay and Moncur, 1996). They are widely used for supporting interactions in the web pages, as they can easily be integrated into many popular browsers. From our study, 12.6% of the collection were found to contain Java applets and 7.7% of the collection were with JavaScript.

Table II shows the changes for the sample web pages. Nearly half of the web pages experienced changes during the monitoring period regardless of the update frequency. These web pages are most likely to contain general news, company product releases, upcoming events, forums, bulletin board, and various promotions. 55.2% of the web page were found without any changes during the monitoring period. Home pages, general information providing pages, company service pages as well as documentation pages about events and facts tend to be static and fall into this category of web pages. It is believed that the higher percentage of web pages with changes could be obtained if longer monitoring period is applied. Of the 105 web pages, 4 pages “disappeared” along the way (i.e. deleted) and no longer exist at the end of the monitoring period. Two pages are from the science domain, one is from the education domain and one is from the society & culture domain.

<table>
<thead>
<tr>
<th>CHANGES</th>
<th>PERCENTAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pages with Changes</td>
<td>44.8%</td>
</tr>
<tr>
<td>Pages without Changes</td>
<td>55.2%</td>
</tr>
<tr>
<td>Dead Pages</td>
<td>3.8%</td>
</tr>
</tbody>
</table>

Changes Based on Web Site Domains

Seven domains were selected from Yahoo!’s directories for the study. They were explored and analyzed to help identify explicit and implicit information that may offer predictive value both for permanence and continuity. Figure 1 shows the average content change percentage by different domains.
As shown in Figure 1, the art domain and the science domain are more stable on content change over time. They are subjected to less change in content than other domains. The average content change percentages for the art and science domains are 5.1% and 2.2% respectively. The education, business, entertainment, society & culture and government domains have quite close content change percentages. This means that a relatively high proportion of information maintained on the web pages is modified when a change is made to these web pages. The average content change percentages for these web pages can reach up to 24.8% for each change.

Figure 2 shows the total changes per page by domains over the one-month period. It can be seen that the education, business and entertainment domains are more active over time and updated more frequently. They have significantly more changes than other domains. The business domain has an average of 9.8 changes per page. The education domain has an average of 9.1 changes per page and the entertainment domain has 7.2 changes per page. In terms of the society & culture and government domains, although average content change percentages are close to the education, business and entertainment domains, they have relatively fewer total changes per page. In addition, it can also be observed that the art domain has the least average content change percentage, but with a higher number of changes per page than the society & culture and government domains. The science domain has only 0.9 total changes per page which is the lowest among all the domains.
Figures 1 and 2 give the general and average update characteristics of web pages by domains. It revealed that the web pages of the education, business and entertainment domains are subject to changes more frequently, and at the same time, they also experience more content changes. Generally, they are good candidates as targeted web pages for monitoring service as they are more active than other domains.

Table III presents more detailed information on updates by domains. The entertainment domain experienced the most change (66.7%) while the science domain experienced the least change (26.7%). In general, the average content change percentage is quite low and most of the web pages have the percentage that is less than 30%. This implies that, generally, not much of the content is changed each time when a page is updated. It can be observed that the pages with high percentage of content change are normally associated with news or documentation pages. In terms of total changes per page, most of the web pages have the total changes that are less than 30 during the monitoring period. Besides the education domain and the business domain, no web pages from others domains are found with total changes that are more than 50 during the monitoring period.

Table III Percentage of web pages by domains

<table>
<thead>
<tr>
<th>DOMAIN</th>
<th>PAGES WITH CHANGES (%)</th>
<th>PAGES (%) WITH AVERAGE CONTENT CHANGE PERCENTAGE</th>
<th>PAGES (% WITH TOTAL CHANGES IN ONE MONTH)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>≤ 10%</td>
<td>&gt;10% &amp; ≤ 30%</td>
</tr>
<tr>
<td>Art</td>
<td>33.3%</td>
<td>26.7%</td>
<td>6.7%</td>
</tr>
<tr>
<td>Education</td>
<td>53.3%</td>
<td>20.0%</td>
<td>20.0%</td>
</tr>
<tr>
<td>Business</td>
<td>60.0%</td>
<td>13.3%</td>
<td>20.0%</td>
</tr>
<tr>
<td>Entertainment</td>
<td>66.7%</td>
<td>6.7%</td>
<td>33.3%</td>
</tr>
<tr>
<td>Science</td>
<td>26.7%</td>
<td>13.3%</td>
<td>6.7%</td>
</tr>
<tr>
<td>Society &amp; Culture</td>
<td>33.3%</td>
<td>13.3%</td>
<td>6.7%</td>
</tr>
<tr>
<td>Government</td>
<td>40.0%</td>
<td>6.7%</td>
<td>26.7%</td>
</tr>
</tbody>
</table>
Domains indicate the nature of the web sites. As the nature of web pages varies from domain to domain, the update behavior differs significantly (Turau 1998). According to the analysis, it becomes possible to understand which domains are more likely to persist or cease to exist and which are more likely to change than others. More specifically, these domains are discussed as follows.

The art domain contains web pages on literature, musical and architectural events, art collections, research perspective in art education and so on. These pages are more focused on presenting the existing facts like art resources and collection gallery. The web pages are updated randomly and the contents that changed in these web pages are mainly on either news releases or images. As expected, images are commonly used in art domain to display photos, pictures and paintings.

The education domain comprises web sites that are primarily used by administrators to disseminate information about the institution and academics to engage in the scholarly communication of their research work and to provide students with online educational learning or teaching materials. These web pages are updated quite frequently and well maintained with a hierarchical structure for readership. However, web designers for academic web sites remain closer to the original purpose of HTML codes with textual-based content.

As a fast growing area, the business domain is one of the most active domains with the current rage of e-commerce and dot.com enterprises. In this highly competitive business world, the web pages in the business domain are updated constantly and frequently to reflect latest developments. These are commercial sites that are mainly used to provide product information and online transaction services. An abundance of animated images, image maps, fancy background images, frames and advanced font handling are characteristics of commercial-related web pages. It is also notable that out of the 15 sample web pages from the business domain, 5 pages provide e-commerce related services. These 5 e-commerce pages have all experienced changes during the monitoring period.

Web pages from the entertainment domain provide a host of entertainment-based information such as movies, entertainment news, activities, CD releases, etc. Multimedia, animated images and visually-rich designs that are hallmarks of the entertainment industry are clearly evident from such web pages. From the study collection, 86.7% and 33.3% of the web pages from this domain is accompanied with images and multimedia respectively. The entertainment domain is found with the highest percentage of web pages with changes. In addition, the average content change percentage of the domain is the highest. The domain also has the most total changes per page. It is quite obvious that the entertainment domain is the most active domain from the sample web pages and it may deserve more attention when they are specified as target pages to be monitored.

The science domain and the society & culture domain are relatively static in terms of total changes per page as well as average content change percentage. The web pages
that changed frequently are mainly those associated with bulletin boards, upcoming events and news updates.

Policies, announcements, documents, news and events usually appear in web pages from the government domain in order to convey information from governments or governmental organisations. The findings indicate that 40.0% of such pages encountered changes during the monitoring period with changes.

Changes Based on Web Page Types

As web pages will be the target for monitoring, this section discusses the changes that occurred in the collected web pages based on web page types. Table IV gives the percentage of web pages that belong to the different web page types. Among them, about 28.6% of web pages are of type Text, followed by 19.1% for Documentation, 12.4% for Database Item, 16.2% for Organisational Page, 9.5% for Multimedia, 8.6% for Tools and 5.7% for Home Page.

<table>
<thead>
<tr>
<th>WEB PAGE TYPES</th>
<th>PERCENTAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organisational Page</td>
<td>16.2%</td>
</tr>
<tr>
<td>Documentation</td>
<td>19.1%</td>
</tr>
<tr>
<td>Text</td>
<td>28.6%</td>
</tr>
<tr>
<td>Home Page</td>
<td>5.7%</td>
</tr>
<tr>
<td>Multimedia</td>
<td>9.5%</td>
</tr>
<tr>
<td>Tools</td>
<td>8.6%</td>
</tr>
<tr>
<td>Database Item</td>
<td>12.4%</td>
</tr>
</tbody>
</table>

Figure 3 shows the average content change percentage by web page type. The average changes for Organisational Page, Documentation and Text types are higher than other types of pages. Web pages of the Text type have the highest average content change of 43.7%. This type of web pages mainly refers to electronic news articles. When they are updated, the average content change percentage is rather high. The content changes for the Organisational Page and Documentation types are 20.2% and 22.7% respectively. The Tools and Home Page types have the lowest average content change of 4.1% and 3.3%. The changes made to these web pages are relatively less compared with other types.
Figure 3 Average content change by web page types

Figure 4 describes the total changes per page by web page type over the monitoring period. Text type achieved the most frequent changes among all the web page types, as web pages of this type are active and updated very frequently. This is followed by the Database Item and Organisational Page types. The total changes per page for the Database Item and Organisational Page are 10.6 and 9.1. In general, the information contained in these web page types also tend to be updated from time to time. Thus, these three web page types, Text, Database Item and Organisational Page, are more active and suitable to be monitored. Although the average content change of the Documentation type is much higher than the Multimedia and Tools types, they have very close total changes per page. These web pages generally provide less active information. As for Home Page type, it provides directories or links to sub-pages as well as general information about organisation, company and personal information. It usually sits at top-level pages such as root pages. Thus, home pages are relatively static.

Figure 4 Total changes per page by web page types
Figures 3 and 4 have shown that web pages of Text type are very suitable to be monitored. They have high change frequency and content change percentage. Organisational Page and Database Item types have relatively high total changes. They are also considered suitable for monitoring purposes. However, Home pages are poor candidates for monitoring.

**Changes Based on Web Page Attributes**

Web pages undergo changes in content over time. The changes can occur in one or more web page attributes. Figure 5 shows the total changes for all the collected web pages over the monitoring period based on web page attributes. The page attributes include text, hyperlink, image, last updated date and layout. The results are interesting. A total of 238 changes were detected for the text change. It is significantly higher than all the other web page attributes. Hyperlink is the next and followed by the last updated date. Layout and image of web pages have not been updated as often as the other attributes. Therefore, text, hyperlink and last updated date are useful attributes for monitoring within web pages. In particular, the text attribute is very important for monitoring purpose as most information are presented in textual content rather than the other attributes.

![Figure 5](image.png)

**Figure 5** Total changes based on web page attributes

Based on the analysis, it is found that the structure of a web page is seldom changed including title, heading, background color, and the arrangement of the attributes. When designing web pages, tables are frequently used to place text, hyperlinks as well as images in a structured manner. Moreover, frames are also commonly used within web pages to separate different sections of a page.

**Change Frequency**

Web page change frequency measures the pace of web pages being updated. Within the given monitoring period, nearly half of the web pages have experienced changes. However, not all the web pages encompass the same change frequency. Figure 6 describes the web page change frequency using a pie chart representation. Due to the monitoring period, change frequencies that are more than one month cannot be detected.
As shown in Figure 6, 31% of the web pages changed weekly and 26% changed daily. This is followed by 15% and 11% for a frequency of 3 days and 14 days respectively. 17% of the web pages were updated randomly or with other frequencies. From these results, a regular frequency of daily and weekly are recommended for monitoring purposes since these are the most likely change frequencies.

![Web Page Change Frequency](image1)

**Figure 6** Web page change frequency

Figure 7 presents the frequencies by domain. Daily frequency is quite commonly used in the entertainment domain and the business domain. The reason is that the information of these domains are fast changing and, therefore, the web pages are also required to be refreshed to reflect the latest improvement and development. This is especially true for those pages providing daily news, weather conditions, or stock information. Weekly frequency is significantly more often applied in the education domain and it is also very frequently used in updating web pages from the business domain and the entertainment domain as well as the government domain.

![Frequency by domains](image2)

**Figure 7** Frequency by domains
Therefore, one basic function that a monitoring system can provide is to let users specify the monitoring frequency such as daily, weekly, etc. The system can then be activated based on the monitoring frequency the users have specified to check the updates of the web pages being monitored. Alternatively, an immediate checking facility can be provided by the monitoring system to enable users to override the defined frequencies of checking in order to carry out an immediate detection of changes to all the web pages been monitored.

Discussion

From the results of the study, we found that text, hyperlink and last updated date are the most important attributes to be monitored for changes. Therefore, monitoring functions like keywords monitoring, hyperlink monitoring and web page updated date monitoring should be supported. In addition, portion monitoring is also very useful to support users to monitor a specific portion of a web page for changes. The portion monitoring can also be extended to monitor the whole web page. These monitoring functions are discussed as follows:

- **Web Page Updated Date Monitoring.** This enables a user to keep track on the last updated date of web pages. After each modification of a web page, the updated date of the web page will be changed. Therefore, the monitoring system just needs to monitor and detect any changes in the updated date of any monitored web pages. This type of monitoring is rather general. It is only useful when users want to keep track of the entire page and just want to be aware of any changes occurred within the web page.

- **Keyword 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 are deleted or new keywords 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, businesses might want to keep track on the latest types of products in the market. For this, they can track the number of links at the products web page. Any changes in the number of the links in the web page will indicated an update of products. As such, the users can be aware of the latest product that has been marketed by the company or older products that might have been phased out. In addition, the monitoring system can also detect whether the hyperlinks of the monitored web page are dead or alive.
• Portion Monitoring. It allows users to monitor a specific portion of a web page for update. A copy-and-paste operation can be used to copy a section of a web page and paste it into a specified area in the monitoring system for subsequent monitoring. The monitoring system will then store the information in its database. When the web page is updated, the system will extract information from the specified portion area and compare it with the stored information for identifying changes. Portion monitoring can be very useful and powerful for monitoring specific information within a web page.

In addition, web page change frequency implies how often the web pages are changed. In order to track web pages properly, appropriate monitoring frequencies should be set accordingly in monitoring systems. As such, it is practical to allow users to specify the monitoring frequency for the web page they have selected. From the study, it showed that the change frequencies of web pages are usually on a daily and weekly basis. Moreover, immediate checking for web page updates can be used to override the predefined frequencies for checking.

Users are allowed to specify web pages for monitoring. They can input the URL of a web page or web site. However, proper guidelines for selecting web pages to monitor should be given. From the study, the education, business and entertainment domains are most active in contrast to other domains. Text pages, Organisational pages and Database pages are also suitable for monitoring. However, this has to couple with the purpose of monitoring by the users.

An Overview of WebMon Web Monitoring System

In association with this study, a web monitoring system, WebMon, has been developed at the School of Computing Engineering, Nanyang Technological University, Singapore. WebMon allows users to specify web pages to monitor and track the web pages on behalf of the user. Once changes on the web pages are detected, the updated information is delivered and displayed to the user automatically to reflect the latest information.

System Architecture

Figure 8 shows the system architecture of the WebMon system that is based on a client-server architecture using Java applets and servlets. The client subsystem supports three major services: User Management Service, Folder Management Service and Monitoring Service. A user interface is designed to support user inputs on specifying monitored web pages, monitoring functions and the frequency to check the updates of a web page. These are developed as Java applet programs that can be handled by conventional web browsers such as Internet Explorer or Netscape Communicator.
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 and check whether any changes occur in the page. The system also contains internal databases for storing the user profiles, personal folders and necessary information of the monitored web page. In addition, a web server is developed to response the client requests.

Monitoring Functions

The WebMon monitoring system provides a friendly interface to support various functions. Figure 9 shows the interface of the system whereby a two-frame structure is used. The left-hand frame with embedded applet provides user management, necessary monitoring functions and the specification of monitoring criteria. In addition, it is also used for folder management that the user’s personal folder with monitored items is
displayed as a tree structure. The right-hand frame is used as a normal browser to display web pages to users. Once the URL is specified in the URL field, the corresponding web page will then be retrieved and displayed. The right-hand frame is also used to display the monitored results.

Figure 9 Interface with monitoring functions of WebMon

The monitoring functions that the system provides include portion monitoring, keywords monitoring, link monitoring, and web page updated date monitoring (date monitoring). The portion monitoring can be extended to monitor the whole web page. Moreover, the user can also specify frequencies for the system to check the updates. The frequencies available in the system are immediate update (Update All Items Now), Daily, 3 Days, Weekly, and 14 Days. Changes in the web pages monitored are automatically sent to the user’s personal folder. Electronic mail is sent to the user as a notification of the changes detected. The user will subsequently access the personal folder to preview the changes.

Conclusion

Users are using the Internet as an important source for various purposes, ranging from pure entertainment, specialized interests, knowledge-sharing, market research to spotting new industry trends. However, web information can be changed and updated at any time. The dynamic nature of web information has posed a serious problem to users to constantly keep track the latest web information in a particular area. This paper uses a sample of 105 web pages over a one-month period to identify and understand the
characteristics of changes and updates of web pages. The results were analysed on the basis of web site domain, web page type, web page attributes and web page change frequency. Such information will prove useful for the design and delivery of web monitoring systems.

A set of functions and features derived from this study are discussed. Monitoring functions, such as web page updated date monitoring, keywords monitoring, link monitoring, and portion monitoring are also described. These forms part of the design phase in the development of WebMon to provide an interactive environment for users to monitor and track web information from different domains over the web. An overview of the system is briefly described.

References

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