Finding Updated Webpages on the Internet

Ben Choi and Xiancong Xiong
Computer Science, Louisiana Tech University, USA
pro@BenChoi.org

Abstract: **For providing timely and accurate information to the web users, this paper proposes methods for search engines to find updated webpages on the Internet. As the contents on the web are rapidly increasing and constantly changing, search engines must employ advanced techniques to keep up with the changes and thus to provide current information for the users. In this paper, we proposed techniques for search engines to update their databases in a timely manner based on the change history, the significance, and the category of webpages. The techniques have been implemented and tested. Our system classified webpages into categories, assigned significance to webpages, and kept records of the frequency of their changes. Then, using our proposed update policies, the system determined when to revisit and update webpages. Our test results show that the proposed techniques are able to keep the contents of the database relatively fresh.**

Index Terms: **Search Engine, Web Technologies, Information Systems, Data Mining.**

I. Introduction

Nowadays most people use search engines to find whatever they need and want. As the contents on the web are rapidly increasing and constantly changing, search engines must employ advanced techniques to keep up with the changes and thus to provide current information for the users. For search engines, the first thing needed to be done is to find as many webpages from the Internet as possible. Search engines constantly try to find new webpages from the Internet by using a process called crawling. The crawling process starts by visiting some known web links (URLs), which are considered as Internet addresses of the webpages or documents. It retrieves new links from those webpages. It then follows those new links and tries to find more webpages. It also saves a lot of information about the webpages in its database, which is used for searching. The process continues as more and more webpages are discovered.

Since the web is evolving, search engines also must constantly revisit existing webpages to keep up with the changes. One research (Ntoulas 2004) showed that 80% of the webpages are not accessible after one year. News websites will constantly update their webpages whenever an event occurs. Online stores will constantly update the prices and the quantities of their products based on the market demands and their inventory. To keep up with the changes, Google, for example, visit over three billion webpages each month (Fetterly et al. 2003).

In this paper, we focus on update policies that address the issues of how often a webpage should be visited and in what order webpages should be visited. One simple policy is to visit all webpages sequentially without any ordering. Since the number of webpages on the internet is at least 7.91 billion as of July 2012 (worldwidewebsize.com), even with the speed of visiting three billion webpages per month, some of the contents may have been outdated for months. The news records in the database may no longer reflect the current events. To address the problem, better update policies must be developed.

In this paper, we proposed update policies based on the change history, the significance, and the category of webpages. The reasoning is that if every time we visit a webpage and find its contents have been changed, then we must visit the page more often. If a webpage is considered more significant, then it should have higher priority to be revisited. If a webpage belongs to a dynamic category, such as news, then it should be revisited constantly.

We developed and implemented a system to test the update policies. The system classified webpages into categories, assigned significance to webpages, and kept records of the frequency of their changes. Then, using the update policies, the system determined when to revisit and update webpages. Test results show that our update policies are able to keep the contents of the database relatively fresh.

The rest of this paper was organized as follows. Section II provided brief background on search engine technologies and outlined works relating to strategies for finding updated webpages on the Internet. Section III outlined our techniques for finding updated webpages. Section IV provided the system implementation used to test the proposed techniques, while section V described testing and outlined the test results. Finally, section VI provided the concluding remarks and the future research directions.

II. Related Research

Finding and updating webpages is only one of many components for building a search engine. After finding a new webpage, a search engine saves a lot of information about the webpage in its database and creates indexes for speeding up keyword searches (Baberwal & Choi 2004). For helping the web users to find the needed information, search engines order the search results and present to the users the most relevant webpages first (Choi & Sumit 2009). Some search engines also arrange webpages into categories and allow the users to focus on the areas of interest (Choi 2001; Choi & Yao 2005, 2008; Peng & Choi 2005; Chen & Choi 2008).

In this paper, we focused on strategies for finding updated webpages on the Internet. One research (Ntoulas 2004) investigated the change rate of webpages and found that 15% of the webpages changed at least once per week. Another

DOI: 02.CIT.2012.2.15

research (Fetterly et al. 2003) found that 66.3% of webpages were between 4KB to 32KB and that larger sized pages changed more often than smaller ones. Cho & Garcia-Molina (2000) found that 20% of the pages changed every day and summarized the rate of change as Poisson model. However, Padmanbhan and Qiu (2000) found that news site (such as MSNBC) did not follow Poisson model and that certain pages was modified repeatedly. Based on statistical models, Coffman et al. (1998) and Wolf et al. (2002) proposed policies for revisiting webpages. Edwards et al. (2001) proposed policies based on the modification history of the webpages that were kept in the database. However, in the highly competitive industry, the techniques used by commercial search engines were not revealed.

III. Techniques for Finding Updated Webpages

In this paper, we proposed techniques for search engines to update their databases in a timely manner based on the change history, the significance, and the category of webpages. We addressed the issues of how often a webpage should be visited and in what order webpages should be visited, by assigning a priority for each webpage. The priority of each webpage is the combination of three factors: modification history, page rank (for significance), and category factor. Each of the factors is associated with its weight. The priority of a webpage is defined as the combination of the three factors and their weights:

$$\begin{aligned} Priority = {} & CategoryWeight \cdot CategoryFactor \\ & + PageRankWeight \cdot PageRankFactor \\ & + ModificationWeight \cdot ModificationFactor \end{aligned}$$

Where, the weights and the factors are all determined through experiments as described in the following sections.

IV. System Implementation

We implemented a system to classify a webpage to a category, to assign page rank (or significance) to a webpage, and to collect the modification history of a webpage. We also used the system to experiment on the effects of the three factors on keeping the contents of the databases fresh. The overview of the system is shown in Figure 1.

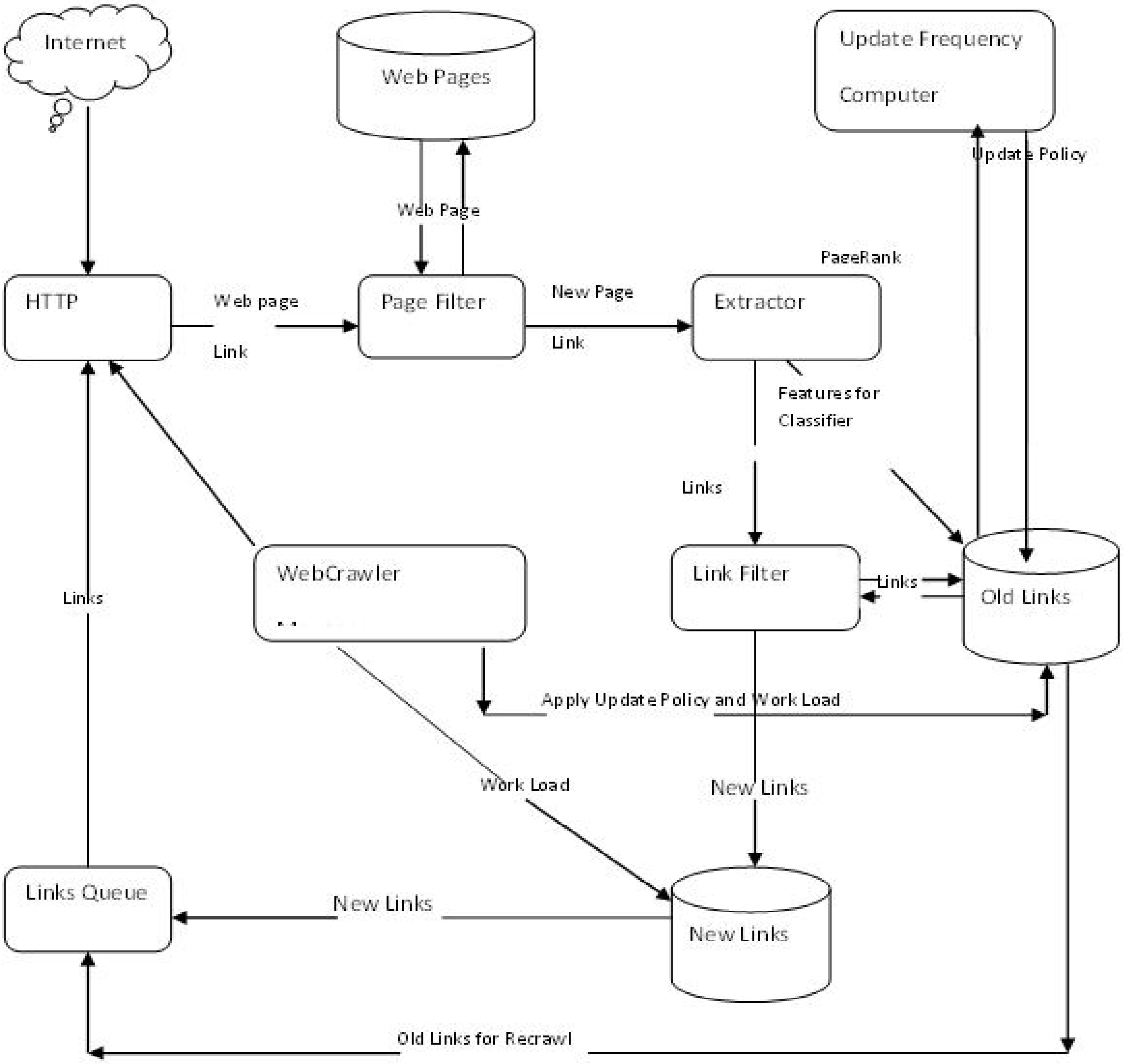
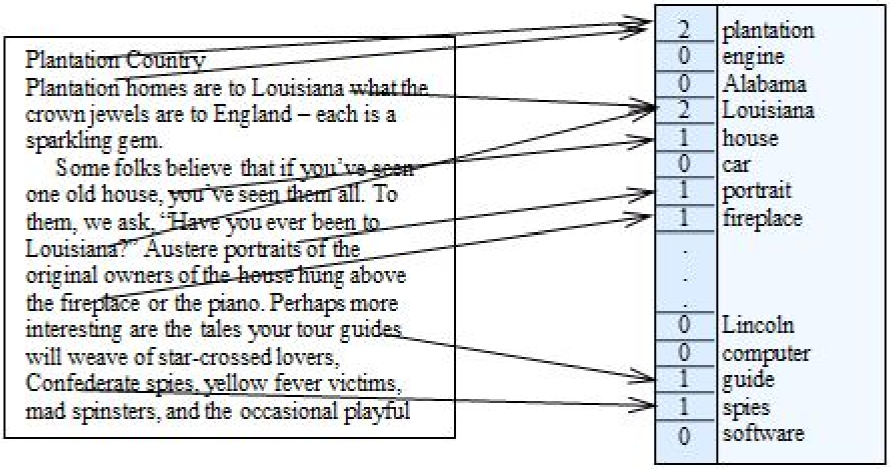


Figure 1. System Implementation for Experiments

DOI: 02.CIT.2012.2.15



Plantation Country
Plantation homes are to Louisiana what the crown jewels are to England – each is a sparkling gem.
Some folks believe that if you've seen one old house, you've seen them all. To them, we ask, "Have you ever been to Louisiana?" Austere portraits of the original owners of the house hung above the fireplace or the piano. Perhaps more interesting are the tales your tour guides will weave of star-crossed lovers, Confederate spies, yellow fever victims, mad spinsters, and the occasional playful

2	plantation
0	engine
0	Alabama
2	Louisiana
1	house
0	car
1	portrait
1	fireplace
. . .	
0	Lincoln
0	computer
1	guide
1	spies
0	software

Figure 2. Representing a webpage by a vector of words

To classify a webpage to a category, we used one of our previously developed automatic webpage classification systems (Chen & Choi 2008). The process to classify webpages begins by analyzing the contents of the webpages. Most automatic classification methods used a vector to represent the contents of a webpage. Figure 2 shows an example of representing the contents of a webpage by a vector of words. Each field of the vector is associated with a word and the number in the field represents the occurrences of that word in the webpage. A category is also defined by a vector, which is obtained by extracting the common characteristics of a group of training webpages used to represent that category. To determine which category a webpage should belong to, the vector of the webpage is compared to the vector of a category to determine their similarity. The webpage is assigned to the category that has the highest similarity (Choi & Yao 2005, 2008). The above classification method uses the text contents of webpages and ignores other features of webpages, such as whether the webpages contain images, phone numbers and prices. Webpage genre classification takes those features into considerations (Chen & Choi 2008). To assign page rank (or significance) to a webpage, we employed the techniques proposed by Page (2001), which considered the number of links referring to the webpage as an important factor. The idea was similar to considering a research paper to be important if it was referred by a large number of other papers. Another factor is the popularity of the webpage, which captures the preferences of millions of users and assumes that if a large number of users like the webpage then it might be more significant (Choi & Sumit 2009).

To collect the modification history of a webpage, we kept a record of change in our database. Each time when our system revisited a webpage, it checked whether the page had been changed or not. To detect the change, a checksum of the page was also kept in the database.

After all the factors were collected, they were weighted and summed to obtain a final priority for a webpage. The webpages having the high priority were revisited first.

V. Testing and Results

Currently, there is no theory that can help determine which factor is more important. We simply designed experiments to try different combinations of the factors and to try to find one combination that can produce the best results.

To measure the test results, we used the Freshness definition (Cho & Garcia-Molina 1999). The Freshness of page P(*i*) at time *t* is defined as:

$$F(P(i);t) = \begin{cases} 1 & \text{If P(i) is up-to-date at time t} \\ 0 & \text{Otherwise} \end{cases}$$

Before trying different combination of factors, we first determined the effect of each individual factor. Figure 3 shows the average freshness of webpage in our test database as the results of using each individual factor to determine the priority of revisiting webpages.

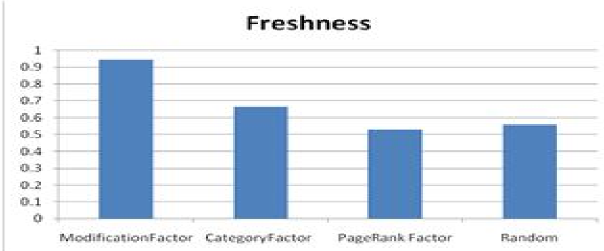


Figure 3. Average Freshness

As shown in Figure 3, the modification factor was important. However, for newly found webpages, there was not sufficient data to determine the modification history. In such case, our test results showed that produced the best results.

$$Priority = 0.5 \cdot CategoryFactor + 0.5 \cdot PageRankFactor$$

By using the equation, the average freshness was 63.6%, which gained 8% improvement comparing to random method. When there were sufficient data to determine the modification history of most webpages in the database, the modification factor became the dominant factor and the category factor became negligible. In such case, our test results showed that

$$Priority = 0.1 \cdot PageRankFactor + 0.9 \cdot ModificationFactor$$

produced the best results. By using the equation, the average freshness was 94.3%, which gained 38.7% improvement comparing to random method.

VI. Conclusion and Future Research

For providing timely and accurate information to the web users, we described techniques for search engines to find updated webpages on the Internet. The techniques made use of the change history, the significance, and the category of the webpages. Based on the test results, we found that the change history was the dominant factor. In care for new webpages, lacking of change history, the combination of the significance and the category factors would improve the results by keeping the contents of the database relatively fresh.

Finding updated webpages on the Internet is only one of many problems that research in search engine technology needs to address. The work reported here in this paper was not complete and much research needs to be done on this area. As the number of webpages on the Internet increases, the problem becomes more urgent.

Future research in search engine technology includes developing completely new search engine. One of the problems of current search engines is that they do not provide sufficient user interactions, in which users simply provide a few keywords, wait for the results, and sequentially scan through pages after pages. A new search engine is being developed that allows users to have more interactions and controls of their Internet experiences (Choi 2006, 2010). The new search engine presented the search results in a hierarchical structure much like a directory tree structure. It not only allowed the users to click on links, but also authorized them to move a link from one directory to another.

Another problem of current search engines and web technologies is that they simply process text contents as patterns without knowing their meanings. The future of information technologies is moving toward semantic web, which aims at automatically extracting useful and meaningful information from the web (Antoniou & van Harmelen, 2004). For a computer to automatically extract useful information from the web, the computer first needs to understand the contents of webpages. This is done with the help of natural language understanding and with the help of assigning meaningful tags to strings of characters. For instance, a string of digits may be assigned as a phone number or a string of digits and letters may be assigned as an address. To keep up with the advanced web technologies, web designers now need to assign meaningful HTML tags to strings, for instance a string of digits may be assigned as "base price" and another as "shipping charge". Understanding web contents will also help organizing webpages into categories (Peng & Choi 2005) and help creating better summary of webpages (Choi & Huang 2009, 2010).

References

[1] Antoniou, G. and van Harmelen, F. (2004). *A Semantic Web Primer*. Cambridge, Massachusetts: The MIT Press.

[2] Baberwal, Sanjay and Choi, Ben. (2004). "Speeding up Keyword Search for Search Engines," *The 3rd IASTED International Conference on Communications, Internet, and Information Technology*, pp. 255-260.

[3] Chen, Guangyu and Choi, Ben. (2008). "Web Page Genre Classification," *The 23rd annual ACM Symposium on Applied Computing*, pp. 2353-2357.

[4] Cho J. and Garcia-Molina H. (1999) "Synchronizing a database to improve freshness". In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 117–128, Dallas, Texas.

[5] Cho, J. and Garcia-Molina, H. (2000) "The evolution of the web and implications for an incremental crawler". In *Proceedings of the Twenty-Sixth VLDB Conference*, pages 200–209, Cairo, Egypt.

[6] Choi, Ben and Huang, Xiaomei. (2009). "Web Page Summarization by Using Concept Hierarchies," *International Conference on Agents and Artificial Intelligence (ICAART 2009)*, pp.281-286.

[7] Choi, Ben and Huang, Xiaomei. (2010). "Creating New Sentences to Summarize Documents," The 10th IASTED International Conference on Artificial Intelligence and Application (AIA 2010), pp. 458-463.

[8] Choi, Ben and Tyagi, Sumit. (2009). "Ranking Web Pages Relevant to Search Keywords," *The IADIS International Conference on WWW/Internet*, pp.200-205.

[9] Choi, Ben and Yao, Zhongmei. (2005). "Web Page Classification," *Foundations and Advances in Data Mining*, chapter: pp. 221 - 274, Springer-Verag.

[10] Choi, Ben and Yao, Zhongmei. (2008). "Web Mining by Automatically Organizing Web Pages into Categories," *Distributed Artificial Intelligence, Agent Technology, and Collaborative Applications*, chapter XII, pp. 214-231, Idea Group Inc (IGI).

[11] Choi, Ben. (2001). "Making Sense of Search Results by Automatic Web-page Classifications," *WebNet 2001 — World Conference on the WWW and Internet*, pp.184-186.

[12] Choi, Ben. (2006). "Method and Apparatus for Individualizing and Updating a Directory of Computer Files", United States Patent # 7,134,082, patent issued on November 7, 2006.

[13] Choi, Ben. (2010). "My Internet," *International Conference on Web Information Systems and Mining*, pp. 171-175.

[14] Coffman, E.; Liu, Jr., Z.; and Weber, R. R. (1998) "Optimal robot scheduling for web search engines". *Journal of Scheduling*, 1(1):15–29, June 1998.

[15] Edwards, Jenny; McCurley, Kevin; and Tomlin, John. (2001) "An Adaptive Model for Optimizing Performance of an Incremental Web Crawler".

[16] Fetterly, D.; Manasse, M.; Najork, M., and Wiener, J. L. (2003). "A large-scale study of the evolution of web pages," *Proceedings of the twelfth WWW Conference*, Budapest.

[17] Ntoulas, Alexandros; Cho, Junghoo; and Olston, Christopher. (2004) "What's New on the Web? The Evolution of the Web from a Search Engine Perspective," *www 2004*.

[18] Padmanabhan, V. N. and Qiu, L. (2000) "The Content and Access Dynamics of a Busy Web Site: Findings and Implications," *ACM SIGCOMM '00 Conference*.

[19] Page, L. (2001). "Method for node ranking in a linked database," US Patent 6,285,999.

[20] Peng, Xiaogang and Choi, Ben. (2005). "Document Classifications Based on Word Semantic Hierarchies," *The IASTED International Conference on Artificial Intelligence and Applications*, pp.362-367.

DOI: 02.CIT.2012.2.15