AN EXPERIMENT ON VISIBLE CHANGES OF WEB PAGES

Joo Yong Lee[1)], Sang Ho Lee[1)], and Yanggon Kim[2)]

Abstract

Since web pages are created, changed, and destroyed constantly, web databases (local collections of web pages) should be updated to maintain web pages up-to-date. In order to effectively keep web databases fresh, a number of studies on the change detection of web pages have been carried out, and various web statistics have been reported in the literature. This paper considers the issues of web page changes in terms of user visuality. First, we consider the effect of a number of tags that do not make difference in terms of user visuality. We learned that approximately 4.5% of web page changes under the byte-wise comparison were unnecessarily determined. Secondly, we investigated the relationship between 'TITLE' tags and 'BODY' tags in terms of web page changes. We found out that an inspection of 'TITLE' tags could allow users to sufficiently determine the change of web pages, so that we can significantly reduce the comparison time of web pages.

1. Introduction

The World Wide Web (WWW) has grown explosively since the early 1990s. According to [10], new web pages are created at the rate of 8% per week. Assuming that the current Web has 11.5 billion pages [5], this result corresponds to 920 million new web pages every week.

As the Web includes most information needed in the world with the rapid growth of the Internet, search engines are becoming the "killer app" of the Web. Whenever users want to look up information on the Web, they get necessary information using a search engine. Search engines such as Google or Yahoo create their own web databases (local collections of web pages) to offer users retrieval service. Remote web pages are changed independently with web databases [2]. Therefore, web databases should be updated to maintain local up-to-date web pages detecting the change of remote web pages. This scenario is illustrated in Figure 1.

[1] School of Computing, Soongsil University, Seoul, Korea. {jylee, shlee}@comp.ssu.ac.kr

[2] Department of Computer and Information Sciences, Towson University, Maryland, USA. ykim@towson.edu

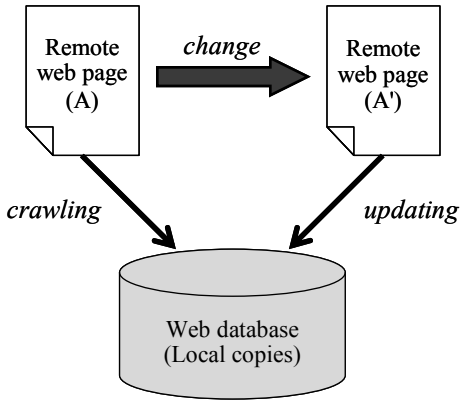

Figure 1. Diagram of updating a web database

A web crawler (also known as a web robot) is a program that automatically collects web pages to create a web database. In general, the web crawler can update its web database in two different ways. Traditionally, the web crawler visits web sites to crawl all web pages and the old collection of web pages is replaced by the newly gathered collection of web pages. We refer to this type of web crawler as a *periodic crawler*. Alternatively, the web crawler can keep already crawled pages after the first crawl and revisit only changed pages (with high probability). The collection of web pages is then partially updated over time. When the web crawler operates in this mode, we call it an *incremental crawler*. The effectiveness of two crawling techniques heavily depends on how often the web page change over time. In principle, if the web crawler can estimate how often web pages change, the incremental crawler is likely to be more effective than the periodic one, in terms of maintaining the "freshness" of web databases.

In order to keep web databases effectively, a number of studies on the change detection of web pages have been carried out. The simplest (but most rigid) way to see if a web page changes is to compare web pages in a byte-by-byte level, which is used in [3, 6, 7]. The byte-wise comparison metric is oversensitive and does not represent the degree of changes. To measure the degree of web page changes, Ntoulas et al. [10] used the TF·IDF cosine distance and the word distance. Lim et al. [9] used a metric based on the edit distance to measure the degree of web page changes. Broder et al. [1] defined the shingling metric and tried to cluster web pages that have the similar contents using the metric, and Fetterly et al. [4] used the shingling metric to investigate how web pages evolve. Kwon et al. [8] showed that each of the existing change metrics often represents the same change of web pages differently, and proposed criteria for web page changes in order to evaluate the effectiveness of the metrics.

Changes of web pages are not always recognized by users who look at web pages through a web browser. Even though web pages are indeed changed, users can regard these pages as the same. Updating web databases for such changes of web pages gives rise to a large amount of processing

overhead; a web crawler requests, downloads, and stores, consuming unnecessary network bandwidth, disk I/Os, disk space, and so on. In our best knowledge, there have been no research activities on web page changes in terms of user visuality.

In this paper, we consider the issues of web page changes in terms of user visuality. We conducted two kinds of experiments. The first experiment shows the effect of a number of tags that do not make difference in terms of user visuality. In the second experiment, we evaluate the change relationship between 'TITLE' tags and 'BODY' tags. From the result, we find out that an inspection of 'TITLE' tags could allow users to sufficiently determine the change of web pages.

This paper is organized as follows. Section 2 shows web page changes that do not make a difference in terms of user visuality and presents experimental results. In section 3, we discuss experimental evaluation of the change relationship between 'TITLE' tags and 'BODY' tags and its results. Section 4 contains the closing remarks and future work.

2. Visible Changes of Web Pages

Changes of web pages are not always recognized by users who look at web pages through a web browser. For example, assuming that a comment is added in the HTML source code of a web page, this change is not noticeable on the web browser and users regard pages before and after the change as the same. We refer to such a change as an *invisible change*. When the change of web pages is not invisible, we call it a *visible change*.

Administrators may think that web databases should be updated to reflect even invisible changes. However, note that the main purpose of web databases of search systems is to offer users retrieval services. Also, updating web databases is expensive. It consumes resources; network bandwidth to download pages, memory to maintain data structures, CPU to evaluate and select URLs, and disk storage to store the text and links of fetched pages. Thus, updating web databases for invisible changes is unnecessary.

2.1. Visible Tags and Visible Attributes

Changes of web pages can be classified into changes of contents, changes of markups, and changes of contents and markups. When web pages change, usually they change in their markups [4]. Figure 2 [4] shows the changes that were additions or deletions of tags, broken down by tags.

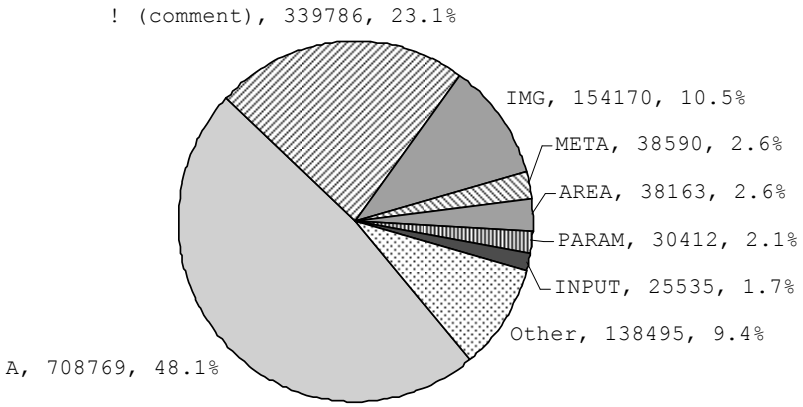

Figure 2. Breakdown of tags that changed

Figure 2 shows that 23% of the changes were within tags that started with the comment character '!', indicating that the change was exclusively to a comment [4]. The '!' (comment) tag, the 'META' tag, the 'PARAM' tag, and so on represent invisible changes of web pages. This surprising result shows that such tags make difference on the change detection under the byte-wise comparison.

Table 1. Visible tags and its visible attributes

Tag	Attribute	Tag	Attribute
<A>	HREF	<OBJECT>	STANDBY
<APPLET>	ALT	<OL>	TYPE
<AREA>	-	<OPTION>	VALUE
<B>	-	<PRE>	-
<BASEFONT>	-	<S>	-
<BIG>	-	<SAMP>	-
<BLOCKQUOTE>	-	<BLOCKQUOTE>	ALT, SRC
<CITE>	-	<STRIKE>	-
<CODE>	-	<SUB>	-
<DD>	-	<SUP>	-
<DFN>	-	<TD>	-
<EM>	-	<TEXTAREA>	-
<FONT>	SIZE	<TH>	-
<I>	-	<TITLE>	-
<IMG>	ALT, SRC	<TT>	-
<KBD>	-	<U>	-
<LI>	TYPE	<VAR>	-

We classify 104 tags [11] into two types: 'visible' and 'invisible' types. This classification is determined by the guides [12] and manual decision. Tags that do not make difference in terms of user visuality are *invisible tags*. Tags that are not invisible tags are *visible tags*. For instance, the '!' (comment) tag is an invisible tag and the 'IMG' tag is a visible tag. Also, we classify attributes of visible tags into two types: 'visible' and 'invisible' types. The 'TARGET' attribute and the 'HREF' attribute of the 'A' tag is an *invisible attribute* and a *visible attribute* respectively. Visible tags and their visible attributes are represented in Table 1.

2.2. Experimental Setup

We examine the aspects of invisible changes of web pages. In particular, we attempted to answer what the effect of tags that do not make difference is, in terms of user visuality, i.e. invisible tags.

Our hardware environment is a machine with 1.8 GHz Pentium IV processor, 1 GB of RAM, and a fast Ethernet network connection. We randomly selected 10,000 Korean sites. We crawled 55,417 pairs of web pages from the sites at intervals of two day. We conducted these crawls using the DSCrawler [7], a web crawler designed to crawl on a seed-by-seed basis.

To answer our question, we process downloaded web pages. The processing removes invisible tags and invisible attributes of visible tags from web pages. We compared the pages under byte-wise comparisons. Figure 3 and 4 show HTML source codes of web pages before and after processing respectively.

2.3. Experimental Results

Table 2 represents the results of comparing web pages. As can be seen, 28.2% of all web pages before the processing were changed, and then 709 (4.5%) pages of those were changed after the processing. In other words, those pages were changed invisibly. From the results, we learned that approximately 4.5% of web page changes under the byte-wise comparison were unnecessarily determined. By avoiding the update for such changes, we can enhance the efficiency of using resources; network bandwidth, memory, CPU, and disk storage.

Table 2. Result of comparing the pages

	No changed pages		Changed pages	
	Number	Percent	Number	Percent
No processing	39,781	71.8%	15,636	28.2%
Processing	40,490	73.1%	14,927	26.9%

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"><html><head><title>♡
nClick='this.style.behavior="url(#default#homepage)";this.setHomePage("http://032we/
24" border="0"></a></td>                                  </tr>              </table
                                                                      <td></td>
d>                            </tr>                                   <tr>
m" method="post" action="/category/search.php" onSubmit="return search_submit();">
g="0" cellpadding="0">                                    <form target="_top" name="to
                        <tr>
dio checked value=1 name=kind onFocus="this.blur()">
>               <tr>                     <td> <table width="100%" border="0" cellspac
><table width="100%" border="0" cellspacing="0" cellpadding="0">
layer"></embed>                          </object></td>          </tr>            <tr>
h/swflash.cab#version=6,0,0,0"  width="220" height="23" id="latest_2" align="">
                        <param name="bgcolor" value="#FFFFFF">
<td width="3" height="3" background="/images/login/login1.gif"></td>
href=/html/company/map.php>location</a></td>
                                         <td width="3" height="3" background="/images
tr align="center" bgcolor="#EFEFEF">
width="9" height="9"> </td>
d>                                       <td background="/images/login/login7.gif"></
="center">                               <tr>                                     <td
                        <td width="50%" align="left"><a href="/html/order/guide.php?
ter"><img src="/images/main/community_2.gif"></td>
ee.com/makeshop/userSourcePerID.php?shop_id=makeshop&id=3195&utilCode=headBannerHigh
table width="100%" border="0" cellspacing="0" cellpadding="0">
                                                        </tr>
                        <td width="3" height="3" background="/images/login/login8.gi
<tr bgcolor="#EFEFEF">
' valign='bottom'> </td>   <td align='left' valign='bottom'><strong> · </strong>
="0"></a></td>
mages/main/phone.gif" width="36" height="36"></td>
eight="8"></td>

</td>

/images/main/title_premium2.gif" width="271" height="25"></td>
="93" height="50" align="center" valign="bottom" bgcolor="#F4F4FF"><a href="/html/re
ulpo2.gif" width="63" height="50" border="0"></a></td>
<tr>                                                            <td height="20" alig
```

Figure 3. HTML source code that not processed

```
<title>♡ Welcome to Inchon Wedding ♡</title>
>
es/main/icon11.gif"></td>                                       <td><font>wedding ho
/html/members/join.php"><font>[member join]</font></a></td>

                              <td></td>
n/bullet_rr.gif"> </td>
nbsp;<font><a href="/html/order/">consultation board</a></font></td>
<td><img src="/images/main/community_2.gif"></td>
                                                                          <td></td>
="/images/main/field5.gif"></a></td>
p;</td>         <td> · <a href='/board/?doc=bbs/gnuboard.php&bo_table=ad&wr_id=14'>Fe
                                                   <td><img src="/images/main/phone.gif
                              <td><font>   shop in shop</font></td>
d>
ef="/html/recommend/index.php?id=6"><img src="/images/best/box_company/best_yaain.gi
<td><font>Chunga jewelry</font></td>                                           <td><f
td>                                                  the formal bows</td>
ion value='wedding-hall buffet'>wedding-hall buffet</option>
                              <td><font>(including both houses)</font></td>

<img src="/images/icon/s_new.gif"> <font>wedding enterprise</font> <font>synchr

s/icon/bt5.gif'> on receive application for an estimate</font></td>    <td><font><im
<td></td>              <td>04-04</td>  <td>wedding-hall buffet</td>     <td>Gyeongg
gif></font></td>              <td></td>              <td>04-04</td>  <td>wedding-
                <td>04-04</td>  <td>wedding-hall buffet</td>     <td>Gyeonggi.Seongn
</td>   <td>Gyeonggi.Bucheon</td>      <td>Mr. Sohn</td>         <td>2005-05-10</td
td>
ttom/banner_naver.gif"></a></td>
/weddingjob.co.kr/"><img src="/images/banner_bottom/banner_job.gif"></a></td>
```

Figure 4. HTML source code that processed

3. Relationship between 'TITLE' and 'BODY'

The change detection of web pages in general is conducted by comparing the full text of web pages. It involves loading the full text of web pages into the memory, consuming quite an amount of memory. However, if the inspection of a specific markup can be used to determine the change of web pages, it may not load the full text of web pages into the memory. This may result in substantial savings in memory and significant improvement in the comparison time of web pages.

3.1. Experimental Setup

In this experiment, we study how strong the relationship is between 'TITLE' tags and 'BODY' tags in terms of web page changes. For our experiment, we used a machine with two 733 MHz Pentium III processors, 1 GB of RAM, and a fast Ethernet network connection. From 50,000 Korean web pages we selected randomly, the DSCrawler downloaded 48,243 pairs of web pages at intervals of two day.

Table 3. Breakdown of web pages that downloaded

Type	Number of pages
'TITLE' tags and 'BODY' tags	38,675
'TITLE' tag	39,203
'BODY' tag	39,599
All	48,243

Table 3 represents the number of pages that include 'TITLE' tags and 'BODY' tags. We used 38,675 web pages that included both of them for this experiment.

3.2. Experimental Results

Table 4 represents the change relationship between 'TITLE' tags and 'BODY' tags. 'TITLE' tags were changed in 3,027 web pages, and in 2,970 (98.1%) pages of those, 'BODY' tags were changed. Also, 'TITLE' tags were not changed in 35,648 web pages, and in 29,816 (83.6%) pages of those, 'BODY' tags were not changed. From the result, we found out that there is the strong relationship between 'TITLE' tags and 'BODY' tags in terms of web page changes.

The body of a web page contains the page's contents. We can think of the body as a canvas where the content appears: text, images, colors, graphics, etc [11]. From this definition, we can regard the

contents of ‘BODY’ tags as the contents of web pages. Thus, we evolve that the inspection of ‘TITLE’ tags can be used to determine the change of web pages. This means that loading ‘TITLE’ tags of web pages into memory is enough for the change detection of web pages, saving substantial memory.

Table 4. Relationship between ‘TITLE’ and ‘BODY’

		‘TITLE’ tag		Total
		Change	No Change	
‘BODY’ tag	Change	2,970	5,832	8,802
	No Change	57	29,816	29,873
Total		3,027	35,648	38,675

We investigate the ending index of ‘TITLE’ tags in web pages. For this purpose, we used 39,203 web pages that included ‘TITLE’ tags. In Figure 5, each dark bar corresponds to the ending index of ‘TITLE’ tags. The light bars show the cumulative distribution.

By examining Figure 5, we can see that 4 kilobytes would be sufficient to hold almost all (99.7%) of the ‘TITLE’ tags. This means that loading 4 kilobytes of web pages into memory is enough for the change detection of web pages. We refer to ‘TITLE’ tags that extracted loading 4 kilobytes of web pages into memory as *4K-‘TITLE’* tags.

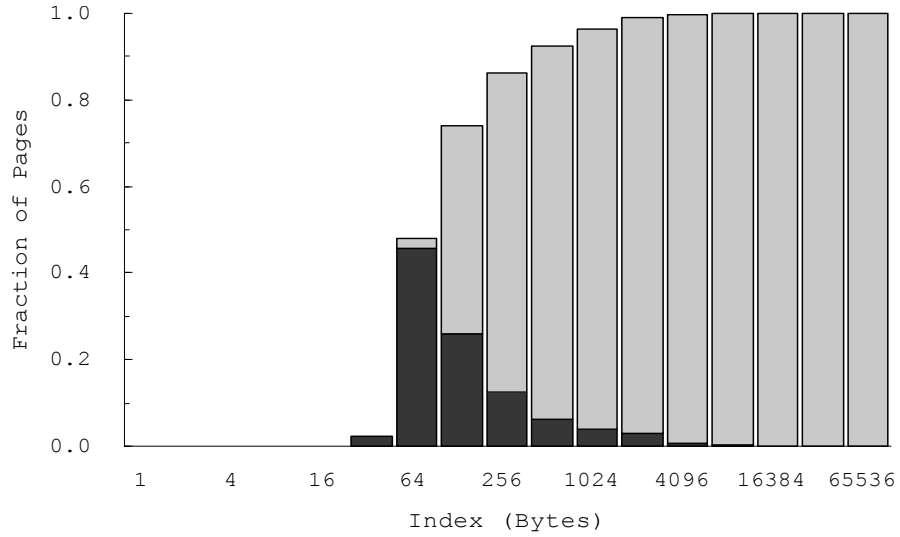

Figure 5. Distribution of the ending index of ‘TITLE’ tags

Table 5 represents the change relationship between 4K-‘TITLE’ tags and ‘BODY’ tags. 4K-‘TITLE’ tags were changed in 3,026 web pages, and in 2,969 (98.1%) pages of those, ‘BODY’ tags were changed. Also, 4K-‘TITLE’ tags were not changed in 35,551 web pages, and in 29,746

Table 5. Relationship between 4K-'TITLE' and 'BODY'

		4K-'TITLE' tag		Total
		Change	No Change	
'BODY' tag	Change	2,969	5,805	8,774
	No Change	57	29,746	29,803
Total		3,026	35,551	38,577

Table 6. Comparison time for the change detection

	'BODY'	'TITLE'	4K-'TITLE'
Time	10h 23m 38s	1h 41m 7s	12m 40s
Ratio	49.2	8.0	1.0

(83.7%) pages of those, 'BODY' tags were not changed. From the results, we found out that the relationship between 4K-'TITLE' tags and 'BODY' tags is strong, in terms of web page changes.

Table 6 represents the comparison time of 'BODY' tags, 'TITLE' tags, and 4K-'TITLE' tags respectively. As can be seen, the change detection using 4K-'TITLE' tags reduces the comparison time by approximately 49 times than using 'BODY' tags. This is a significant improvement in terms of the time efficiency.

4. Conclusions and Future Work

We have considered the issues of web page changes in terms of user visuality. First, we showed the effects of invisible tags, and learned that updating web databases for invisible changes is unnecessary. Secondly, we presented that an inspection of 'TITLE' tags can be used to determine the change of web pages. This results in substantial savings in memory and significant improvement in the comparison time of web pages.

As the Web is going to expand , updating web databases will become more and more difficult. We believe that our findings will be good guides in order to effectively keep web databases fresh.

The term "false negative" is referred for a process that determines changed web pages to be unchanged. The above change detection using 'TITLE' tags may cause occurrences of false negatives. We need to have a way to reduce false negatives. How to effectively reduce false negatives remains as future work.

Acknowledgments

This work was supported by Korea Research Foundation Grant (KRF-2006-005-J03803).

References

[1] Broder, A.Z., Glassman, S.C., Manasse, M.S., and Zweig, G., Syntactic Clustering of the Web, Computer Networks and ISDN Systems, Vol. 29, No. 8-13, pp. 1157-1166, 1997.

[2] Cho, J. and Garcia-Molina, H., Synchronizing a database to Improve Freshness, Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 117-128, 2000.

[3] Cho, J. and Garcia-Molina, H., The Evolution of the Web and Implications for an Incremental Crawler, Proceedings of the 26th International Conference on Very Large Data Bases, pp. 200-209, 2000.

[4] Fetterly, D., Manasse, M., Najork, M., and Wiener, J.L., A large-scale study of the evolution of Web pages, Software – Practice and Experience, Vol. 34, No. 2, pp. 213-237, 2004.

[5] Gulli, A. and Signorini, A., The Indexable Web is More than 11.5 billion pages, Proceedings of the 14th International Conference on World Wide Web, pp. 902-903, 2005.

[6] Kim, S.J. and Lee, S.H., An Empirical Study on the Change of Web Pages, Proceedings of the 7th Asia-Pacific Web Conference, pp. 632-642, 2005.

[7] Kim, S.J. and Lee, S.H., Implementation of a Web Robot and Statistics on the Korean Web, Proceedings of the 2nd International Conference on Human.Society@Internet, pp. 341-350, 2003.

[8] Kwon, S.Y., Lee, S.H., and Kim, S.J., A Precise Metric for Measuring How Much Web Pages Change, Proceedings of the 11th International Conference on Database Systems for Advanced Applications, pp. 557-571, 2006.

[9] Lim, L., Wang, M., Padmanabhan, S., Vitter, J.S., and Agarwal, R., Characterizing Web Document Change, Proceedings of the 2nd International Conference on Web-Age Information Management, pp. 133-144, 2001.

[10] Ntoulas, A., Cho, J., and Olston, C., What's New on the Web? The Evolution of the Web from a Search Engine Perspective, Proceedings of the 13th International Conference on World Wide Web, pp. 1-12, 2004.

[11] Raggett, D., HTML 4.01 Specification, http://www.w3.org/TR/1999/REC-html401-19991224, 1999.

[12] Treviranus, J., Authoring Tool Accessibility Guidelines 1.0, http://www.w3.org/TR/2000/REC-ATAG10-20000203, 2000.