©Journal of the Technical University Sofia, branch Plovdiv
“Fundamental Sciences and Applications”, Vol. 16, 2011
International Conference Engineering, Technologies and Systems
TechSys 2011
BULGARIA

ANALYZING OF THE EVOLUTION OF WEB PAGES BY USING A DOMAIN BASED WEB CRAWLER

ERDINC UZUN, TARIK YERLIKAYA, MELTEM KURT

Abstract. *To improve algorithms that are used in search engines, crawlers and indexers, the evolution of web pages should be examined. For this purpose, we developed a domain based crawler, namely SET Crawler, which collects the web archives between 1998 and 2008 of three Turkish daily popular newspapers (Hurriyet, Milliyet and Sabah). After completion of the crawl, we obtained a set of 3430997 HTML pages. While the average file size of one web page in 1998 approximately is 5.19 KB, this size in 2008 is 53.94 KB. When considering the size of main contents of web pages are similar, this observation shows the degree of increase in the use of unnecessary contents and tags. Analyses indicate that the use of link, image and layout tags has increased significantly in the last decades. Moreover, the <div> tag has been used instead of the <table> tag, especially in Milliyet and Sabah.*

Key words: *Web Crawlers, Web Evolution, Degree of Changes in Web Pages*

1. Introduction

The web is open data sources for researches on subjects of information retrieval, natural language processing and data mining. The growth of these data sources has been exponential. However, the use of redundant contents such as advertisements, banners, navigation panels and comments in web has been exponential, too. Web archives can be examined for better understanding of “exponential growth”. In this study, we collected the web archives of Turkish online newspapers: Hurriyet (http://www.hurriyet.com.tr), Milliyet (http://www.milliyet.com.tr) and Sabah (http://www.sabah.com.tr).

For preparing this collection, we developed a domain based crawler that is namely SET Crawler. (SET (Search Engine for Turkish) project has a search engine, an evaluator module, and a crawler module. Classes of SET, are open source, are available via the web page http://bilgmuh.nku.edu.tr/SET/.) This crawler firstly produces a list of URLs, this list is called Seeds. Then, as it visits these seeds, it gets all hyperlinks in the visited pages and appends them to the list of URLs. While this crawler extracts hyperlinks, it also applies URL normalization that is a processing of transforming URL strings into canonical form.

Fetterly et al. [1] collected 720000 web pages drawn from 270 web servers on a daily basis over the course of 4 months. It expands on researches of Cho and Garcia-Molina[2]. They discovered that about 40% of all Web pages altered within a week and 50% of them changed within 50 days. They also found that 50% of these pages in “.com” changed within 11 days. On the other hand, it took 4 months for half of the “.gov” pages to have altered. Sun et al. [3] compared the length of the pages and the number and lengths of embedded images and frames by using a set of 100000 web pages from the Open Directory listings. Their examination shows that 40% pages changes signatures. In this study, we examine the changes in file size and tags (<div> and <table>) used in layouts per year. For these examinations, we crawled a set of 3430997 web pages between 1998 and 2008 for each newspaper.

The composition of this paper is as follows. The next section (Section 2) describes the SET Crawler that is used for collecting archives of web pages. Section 3 compares the average of file sizes and the number of layout tags of each newspaper per year. Section 4 concludes this paper.

2. SET Crawler

The SET Crawler is designed for collecting archive of three major Turkish daily newspapers as Hurriyet, Milliyet and Sabah. These newspapers contain news about politics, economic reviews, international, culture and sports.

The SET crawler firstly generates the URL listing by producing an appropriate date form. These generated listings are seeds for our crawler. For example:

- http://www.milliyet.com.tr/2005/01/08/
- http://hurarsiv.hurriyet.com.tr/goster/haberler.aspx?id=1&tarih=2007-04-19
- http://arsiv.sabah.com.tr/2009/02/01/

Then, it downloads initial URL, identifies all the hyperlinks in the downloaded page and adds them to the URL Listing. Before adding URL, URL normalization process and unique web page control process are applied to URL for ensuring a duplicate record is not created.

Algorithm 1
Preparing of the URL Listing

```
for all days in a given year do
 generate hyperlink[day]
end for
for all hyperlinks do
 download a given page and resolve URLs in a downloaded page
 URL normalization process for resolved URLs
 if URL contain web domain and date value of a generated domain then
  if URL contain unique page id then
   if hash table do not contain page id then
    add page id to hash table
   add hyperlink listing
   end if
  end if
  else
  add hyperlink listing
  end else
 end if
end for
```

URL is a string representing a web page. The URL contains a protocol, authority, path, query and fragment components [4]. The protocol, *http://* in our study, is used for transferring data between a web server and a client. The authority has three additional information such as host, user and port. The path contains directories including a web page or a file name. The query string, starts with the question symbol "?", has two parameters that are names and values. The fragment allows indirect identification of a secondary resource by reference to a primary resource and additional identifying information. The following example shows an URL from Hurriyet.

$$\underbrace{\text{http://}}_{\text{Protocol}}\underbrace{\text{hurarsiv.hurriyet.com.tr}}_{\text{Authority}}\underbrace{\text{/goster/haberler.asp}}_{\text{Path}}$$

$$\text{...aspx?}\underbrace{\text{id=1\&tarih=2007-04-19}}_{\text{Query}}\text{\#}\underbrace{\text{konu1}}_{\text{Fragment}}$$

Crawlers employ URL normalization in order to avoid adding same pages to their URL Listings. There are several types of normalization such as converting the URL to lowercase, adding trailing "/" and removing the fragment, dot-segments and arbitrary query string variables. Our crawler has two additional parts in addition to URL normalization.

The first part is designed for an issue in the links of web pages that contain only a filename in hyperlink. Therefore, our crawler automatically adds the information of protocol, authority and path by determining which domain and directory are in URL. For example, a web page contains a filename "akbal.html" in its hyperlink. Our crawler transforms this filename to

http://www.milliyet.com.tr/1997/01/14/yazar/akbal.html

Another part is designed for avoiding duplication in web pages. For instance:

hurarsiv.hurriyet.com.tr/goster/haber.aspx?id=5705456

The query has two names as "id" and "tarih (date)". This web page can be accessible from different URLs by chancing date value so the identifier value is "id" in this query. Therefore, our crawler creates an additional control list for unique parts of URLs to avoid same web pages in URL Listing.

In implementation of this crawler, we used C# programming language that is designed to work with .NET platform. In the downloading phase, SET Crawler downloads web pages and saves them to a local disk by importing the namespace System.Net and using the classes WebRequest and WebResponse. In the construction of URL Listing, Regular expressions and hash table are used. A regular expression is special text string for describing a search pattern. Moreover, regular expressions are used by many programming languages such as C, C++, .NET, Java, Perl, Ruby, Python to search and manipulate text based on patterns. For accessing .Net regular expression engine from C#, we import the namespace

"System.Text.RegularExpressions".

Hash table is a data structure that uses a hash function to map unique keys to their associated values. In searching of links, hash tables are good for a quick search on the URL Listings. Hashtable class in C# represents a hash table. Before using Hashtable class, we import the namespace “System.Collections”.

3. Experiments

In experiments, 3430997 web pages are crawled for each newspaper between 1998 and 2008. The total size of web pages is around 109 GB in UTF-8 format. The average size of a web page is approximately 33.35 KB. Now, we give more information about this collection. Fig. 1. indicates the count of downloaded web pages for each newspaper per year.

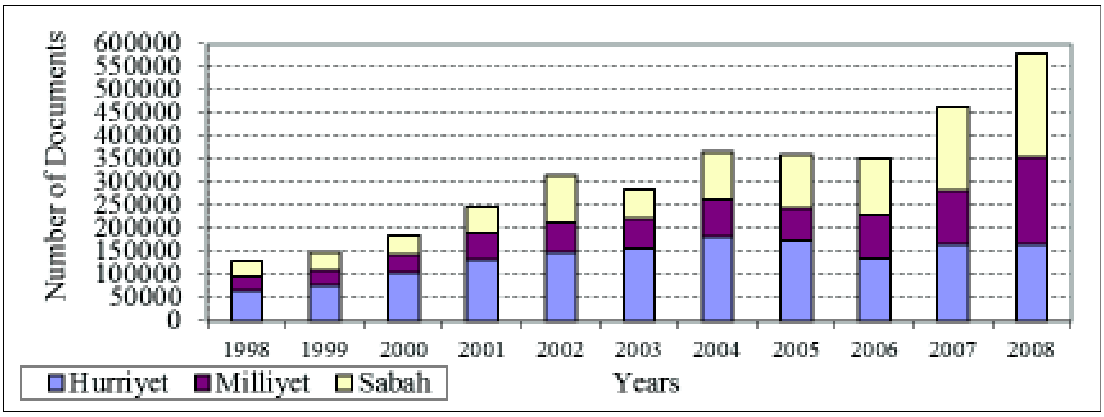

***Fig. 1.** The number of crawled documents per year*

The numbers of crawled web pages in 1998s are 64468, 32333, 31482 for Milliyet, Hurriyet and Sabah, respectively. The total number of this year is 128283. In 2008s, the numbers of obtained web documents are 165872, 188909 and 233992 for newspapers. The total number of crawled pages is 578773. That is, the total number of crawled web pages in 2008s is roughly 4.5 times larger than the total number of web pages in 1998s. In some years, the number of obtained web pages decreases because the web sites try to use different web design software. However, still the number of web pages has increased. As a result, we crawled approximately 1508975, 836375, 1085647 web pages for Milliyet, Hurriyet and Sabah, respectively. Now, we examine the total file sizes of obtained web pages. (See Table 1.)

Table 1. The Total sizes of crawled web pages per year

	MB			
Year	**Hurriyet**	**Milliyet**	**Sabah**	**Total**
1998	395	117	118	650
1999	552	317	541	1049
2000	1134	521	666	2321
2001	2487	1018	1293	4799
2002	3822	1244	1001	6067
2003	4998	930	799	6728
2004	6008	1240	5393	12642
2005	4563	1043	6307	11913
2006	3508	2084	7239	12831
2007	7934	1763	12210	21907
2008	7923	6429	16136	30488
Total	43324	16707	51725	111756

In Table1., the total size of crawled web documents in 1998s is approximately 650 MB. However, the total size increases to 30488 MB in 2008s. In the other words, the file size of downloaded pages in 2008s is roughly 46.9 times larger than the count of web pages in 1998s. The following table (Table 2.) shows the average size of a web page for three newspapers per year.

Table 2. The average sizes of a web page for each newspaper

	KB			
Year	**Hurriyet**	**Milliyet**	**Sabah**	**Avg.**
1998	6.27	3.71	4.50	5.19
1999	7.33	10.07	13.64	9.93
2000	11.16	13.42	16.64	12.86
2001	19.09	18.16	24.42	20.05
2002	26.32	19.74	10.01	19.68
2003	32.48	15.03	12.66	24.12
2004	33.56	16.24	53.15	35.42
2005	26.77	15.83	54.28	33.79
2006	26.67	22.36	60.45	37.24
2007	49.17	15.48	68.89	48.39
2008	48.91	34.85	73.77	53.94
Avg.	29.40	20.45	48.79	33.35

In Table 2., the average file size of a web page in 2008s is roughly 10.4 times larger than the average size of a web page in 1998s. The file size of main content is approximately 2 KB so the other parts of a web page contain tags and unnecessary contents. In 2008, the average file size of a web page is 53.94 KB. Approximately 51 KB of a web page includes tags and redundant contents. When viewing web pages per year, the use of redundant contents such as advertisements, banners, navigation panels and comments in web pages has been exponential has increased exponentially. For example, there were no advertisements in old years. However, the number of advertisements increased for every year. In the last experiments, we examine important tags such as link, images and layout used to create web pages.

In this examination, we calculate the use of these tags in web pages. For this calculation, regular expressions can be used. Table 3. indicates patterns that is used for calculating of tags.

Table 3. Patterns used in SET Crawler

	Patterns
Links	(?<=href=\").*?(?=\") openWindow\('(.*?)'\) popup\('(.*?)'
Images	<\\s*img [^\\>]*src\\s*=\\s*([\"\\'])(.*?)>
Table	</table>
Div	</div>

HREF indicates the URL being linked to. Our examination of tags in web pages shows that javascript functions such as openWindow and popup in web pages are also used the URL being linked to. Therefore, we use three regular expression patterns for extracting links from web pages. Table 4. gives the average number of links obtained from web pages per year.

Table 4. The average number of Links for each newspaper

	Hurriyet	**Milliyet**	**Sabah**
1998	17.31	1.47	3.37
1999	20.76	23.10	3.74
2000	26.30	31.62	39.54
2001	34.51	46.81	39.77
2002	73.86	48.88	64.04
2003	77.12	45.30	72.46
2004	80.60	54.01	76.95
2005	79.28	53.68	93.89
2006	26.43	47.80	104.25
2007	57.46	48.91	134.28
2008	76.91	50.12	150.65

Table 4. indicates that the average number of links frequently increased every years. In 1998 of Milliyet and Sabah and 1999 of Sabah, the average numbers of links are too low because the <frame> tag is used in these years and links are created in one frame. The <frame> tag defines one particular window within a frameset. The other examination is the average number of images used in web pages. (See Table 4.)

Table 5. The average number of Images for each newspaper

	Hurriyet	**Milliyet**	**Sabah**
1998	3.93	1.82	3.06
1999	5.64	3.37	3.12
2000	5.76	4.00	12.89
2001	6.33	45.66	11.99
2002	113.29	48.22	69.20
2003	122.16	44.17	139.91
2004	126.06	43.71	153.62
2005	126.95	47.71	167.29
2006	21.38	42.51	181.55
2007	81.39	42.63	165.57
2008	113.35	19.53	136.59

In old years, texts and links were used the design of web pages. However, the use of images increased after 2000. In these years, images utilized in menus, links, content of news and advertisements.

The last examination is the average number of <div> and <table> tags obtained from web pages. Table layouts first began appearing in 1993[5]. Div tags were accepted by the World Wide Consortium (W3C) in 1997 [6]. These tags are useful for web designers and researches. For example, Yerlikaya and Uzun developed an intelligent browser that can be viewed only relevant contents to the users by using <div> and <table> tags. Recently, the <div> tag are considered as more suitable method for designing web layouts by most web designers because of the Cascading Style Sheets (CSS) that allows designers to change the entire look of a web site with the use of one or more external style sheets. Figure 2. shows the use of these tags per year. In 1998s, <frame> tag were used. Between 1999-2007, web designers of newspapers mostly utilize <table> tag. After 2007, Milliyet and Sabah designers preferred to use <div> tags for utilizing the advantage of CSS.

5. Conclusion

In this study, we describe the SET Crawler and examine the text collection obtained with this crawler. These examinations can be useful information to improve methods that are used which use web pages as data sources, especially in information retrieval, natural language processing and data mining.

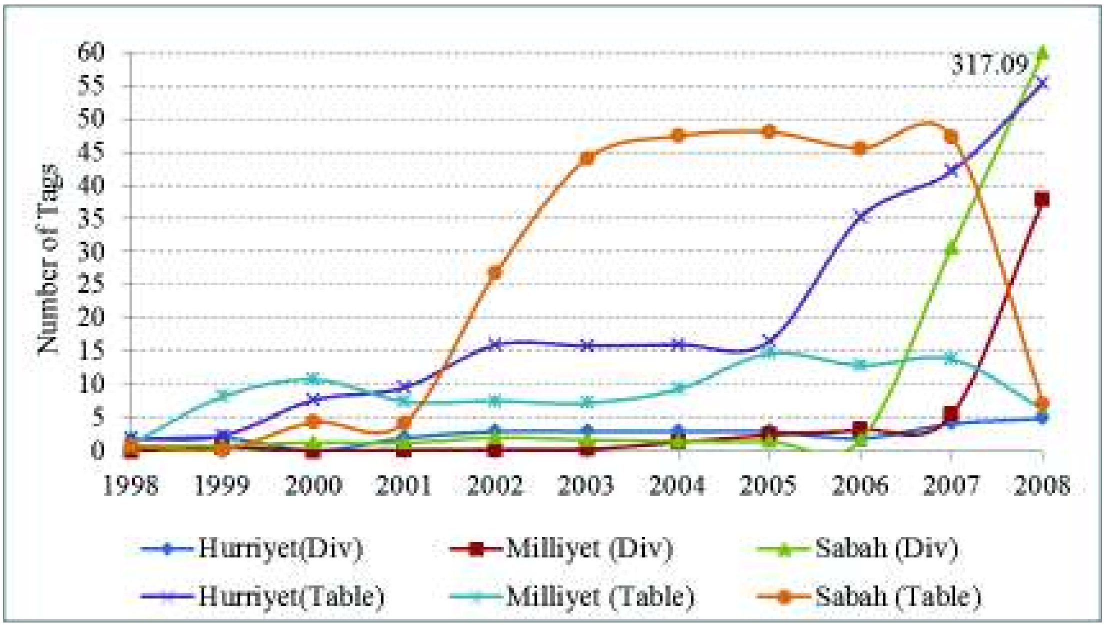

Fig. 2. *The average number of layout tags that is used in web pages*

Fetterly et al. [1] argue that their statistical observations of the measurements showed that page size was strong predictor of both frequency and degree of change. Similarly, our examinations support this claim. The average page size of 2008s is approximately 10.4 times larger than the page size of 1998s.

Some future research possibilities are adapting this crawler to download different web sites and examine these sites. Furthermore, another future work would be to utilize these results to improve methods that are used in our studies such as content extraction, search engine and duplicate document detection.

References

1. Fetterly D., M. Manasse, M . Najork, J. Wiener A large-scale study of the evolution of Web pages. Proceedings of the 12th International World Wide Web Conference, May 2003. ACM Press, 2003; 669–678.

2. Cho J., H. Garcia-Molina The evolution of the Web and implications for an incremental crawler. Proceedings of the 26th International Conference on Very Large Databases, September 2000. Morgan Kaufmann, 2000; 200–209.

3. Sun Q., D. Simon, Y. Wang, W. Russell, V. Padmanabhan, L. Qiu Statistical identification of encrypted Web browsing traffic. Proceedings IEEE Symposium on Security and Privacy, May 2002. IEEE Computer Society, 2002; 19–30.

4. Berners-Lee T. Uniform Resource Identifier (URI): Generic Syntax, 2005; http://tools.ietf.org/html/rfc3986.

5. Raggett D. HTML 3.2 Reference Specification. *W3C*. W3C, 14 Jan. 1997.

6. Wroblewski L. *Site-Seeing: A Visual Approach to Web Usability*. New York: Hungry Minds, 2002.

7. Yerlikaya T., Uzun E İnternet Sayfalarında Asıl İçeriği Gösterebilen Akıllı Bir Tarayıcı. *Akıllı Sistemlerde Yenilikler ve Uygulamaları Sempozyumu (ASYU-2010) 2010*; 21-24 Haziran, Kayseri & Kapadokya, ISBN: 978-975-6478-60-8, 53-57.

8. Uzun E Html İçinde Gereksiz Kelimeleri Çıkaran Benzer Metin Tespit Uygulaması. *Akıllı Sistemlerde Yenilikler Ve Uygulamaları Sempozyumu (ASYU-2010) 2010*; 21-24 Haziran, Kayseri & Kapadokya, ISBN: 978-975-6478-60-8, 48-52.

Department of Computer Engineering, Namik Kemal University, Corlu Engineering Faculty, Corlu / Tekirdag / Turkey
E-mail: erdincuzun@nku.edu.tr

Department of Computer Engineering, Trakya University, Ahmet Karadeniz Yerleskesi, Edirne / Turkey
E-mail: tarikyer@trakya.edu.tr
E-mail: meltemkurt@trakya.edu.tr