Web Forums Change Analysis

Tomasz Kaczmarek and Dawid Grzegorz Węckowski
Department of Information Systems, Poznań University of Economics, Poznań, Poland

Keywords: Web Crawler, Web Forum, Web Page Changes.

Abstract: In this paper we present results from an experiment conducted on over 27 900 web pages gathered every 2 hours over 22 days from 16 forums (4256 independent crawls), to investigate how these web pages evolve over time. The results of the experiment became a basis for design choices for a focused incremental crawler, that will be specialized for efficient gathering of documents from web forums, maintaining high freshness of the local collection of obtained pages. The data analysis shows, that forums differ from generic web portals and identifying places in the source navigational structure, where new documents occur more often, would allow to improve the crawler's performance and the collection freshness.

1 INTRODUCTION

This article reports on the results of experiment conducted on 16 web forums crawled for web pages. The experiments were conducted to gather data concerning web page changes. On one hand we would like to validate the results obtained previously in similar experiments, but for a slightly different conditions (more general web sites and less granular data). On the other hand we are looking for the clues about the possible patterns present in web page change data.

Our main motivation for analysing the data is to improve over existing results in web crawler design and in particular approaches to scheduling web pages for revisitation in subsequent (or incremental) crawls. The goal for crawling would be to capture new information published on the forum as early as possible, without overburdening the crawler infrastructure and the target websites with unnecessary fetches.

2 RELATED WORK

The work related to our concerns several areas that have much in common: research on the Web structure and its changes, methods of change detection, incremental crawler design issues and websites navigation issues including crawling specific types of websites (focused crawling).

2.1 Web Structure and Changes

We base our work on seminal papers by Cho and Garcia-Molina (Cho and Garcia-Molina, 2000), where they made similar investigation of web page evolution in order to design a better incremental crawler. They conducted their analysis on a large collection of web pages (over 500 000) and conclude that pages change on average every 10 days (with 1 day granularity of measurement) with significant differences between commercial and less business oriented pages. Secondly, they found that web pages' changes follow the mathematical model of Poisson process, which is used for sequences of random events happening independently with fixed rate over time.

The same authors continue their work to refine the mathematical model for estimating frequency of change and optimize the access time for incremental crawler in (Cho and Garcia-Molina, 2003). Their experiments show, that using the right estimates the performance of the web crawler can be significantly improved, resulting in more changes being detected.

Further work on exploiting knowledge of the web page changes characteristics focuses on optimisation of the crawler. The work of Baeza-Yates and colleagues from 2005 (Baeza-Yates et al., 2005) discusses strategies for prioritizing web page download. They focus on strategies that take into account web page importance (calculated mainly using Pagerank). They concluded, that using the right strategy it is possible to find the high quality pages sooner, than with breadth-first strategy for example.

Kaczmarek T. and Węckowski D..
Web Forums Change Analysis.
DOI: 10.5220/0004373201050110
In *Proceedings of the 9th International Conference on Web Information Systems and Technologies* (WEBIST-2013), pages 105-110
ISBN: 978-989-8565-54-9

Another analysis of web page change evolution patterns was presented in (Adar et al., 2009b). The analysis was conducted on a finer grain than the whole-page comparison because structural elements of web pages were analysed. Each of 55 000 URLs sampled based on user visits in this study was crawled on a hourly basis which allowed also for detailed analysis of the time of changes. The findings of this study are that the rates of change were higher than in previously reported experiments (averages in the range of tens of hours), with a large portion of pages changing more than hourly. Detailed content and structure analyses identified stable and dynamic content within each page. Using the same data set the Authors reported also in (Adar et al., 2009a) the discovered resonance between user behaviour patterns (frequency of visits or returns to the page) and the page change patterns.

In (Ben Saad and Gançarski, 2011) the Authors took another approach and tried to identify periodic patterns of change on the monitored web pages, in order for the crawler to be able to predict the exact time when the change is most likely to occur on a given URL. The research is based on observations that for some frequently updated web pages (several times per day) the periodic activity can be detected. The approach was applied for web archiving to improve the quality of the web archive.

2.2 Methods of Change Detection

The method of change detection can be crucial for crawling web forums and adjusting revisiting strategy for keeping the local collection of web pages copies up-to-date. One of the simplest solution is the use the document digest to compare subsequent versions of a page. This method is efficient, however the disadvantage of such approach is that both small and large changes are equal. The fact of modification can be also established upon the information in HTTP header `Last-Modified`, although in case of dynamically generated web pages the information can be unreliable. Several other solutions were developed for more robust change detection and description. One of the first solutions designed for HTML documents comparison is HtmlDiff program, developed at AT&T Bell Laboratories (Douglis and Ball, 1996; Douglis et al., 1998). It takes advantage of the algorithm proposed by Hirshberg to solve the longest common subsequence (LCS) problem (Hirschberg, 1977). Rocco, Buttler and Liu (Rocco et al., 2003; Buttler et al., 2004) introduced mechanism called Page Digest for storing and processing web documents. Page Digest introduces a clean separation of the content of a web page and the page structure, and can operate in linear time, with 75% improvement as compared to other solutions. The analysis of well known document similarity metrics (byte-wise comparison, TF-IDF cosine distance, word distance, edit distance, and w-shingling) was also performed in (Kwon et al., 2006). On a structural level of HTML documents (Adar et al., 2009b) presented algorithms for describing DOM tree elements modifications as well as persistence of structural blocks. Several recent studies make use of visual features of web pages to detect and measure changes (Saad and Gançarski, 2010; Law et al., 2012).

2.3 Incremental Crawler Design

The work by Cho on web page evolution, as mentioned earlier (Cho and Garcia-Molina, 2000), allowed them to formulate certain clues for incremental crawler design. Significant changes in the rate of change between different pages (large number of them did not change over the whole 4 month period of the experiment) suggests that crawling should be guided by the rate of change of the web page, not to waste the resources on pages that change less frequently. Secondly, the Poisson process model for web page changes suggests a strategy for incremental crawling, which depends on estimation of which pages are likely to change in a given time period according to the process model and history of changes, and scheduling them first for crawling.

A number of other approaches to prioritize pages in the crawl queue were proposed, that take into account factors other than the frequency of changes of the web page. As (Liu et al., 2011) reports, measures of page importance stemming from web graph structure (like in-degree or PageRank) were successfully evaluated. Web page topic, web site structure or even direct mining of user interest from the search engine log were also proposed as factors to take into consideration when crawling the Web.

2.4 Focused Crawling

There were several efforts aiming at development of methods and building web crawlers, focusing on retrieval of specific web pages.

A focused crawling for collecting mainly novel pages can be established with the use of novelty measure proposed in (Toyoda and Kitsuregawa, 2006). It can be used to find emergent information according to some topical range, e.g. a user's interest. A focused crawler can use the distribution of the novelty measure calculated for current local collection of web

pages, to adjust revisitation strategy, and explore the areas of the Web that most probably contain novelties.

(Baeza-Yates and Castillo, 2007) discussed the most common ways of defining the range of pages to be retrieved from the infinite Web. One can download purely static pages or use only one set of parameters in the URL, alternatively pages can be collected until specified total / per domain amount or until reaching certain number of levels per web site. Authors propose various models of random surfing in a generic unbounded web site and they show the main areas of focus for a web crawler by analyzing how deep the users generally go into the web site.

Recently many solutions deal with web page retrieval from web forums. (Cai et al., 2008) proposed an intelligent forum crawler, called iRobot, which is able to choose traversal paths for visiting different types of pages, by analysing content and structure of a forum web site. The crawler tries to reconstruct the forum's sitemap, with the use of preliminary sampling of web pages from the site and grouping them according to content layout characteristics. The sitemap allows iRobot to select an optimal path for web page retrieval, including only informative pages, discarding duplicates and invalid ones.

(Yang et al., 2009) studied incremental crawling of web forums. They showed that traditional approaches to crawling web pages may be insufficient in terms of forums sites, which have different characteristics than general web sites and contain interrelationships between pages. Determining revisiting strategy based on individual pages change frequency can be inefficient, because within web forums each list of threads or list of posts is split among several pages, implying that one change in a list can cause modification of all related pages. Thus, a list-wise crawling strategy was proposed, after reconstructing a forum linking structure, posts, that are spread across different pages but are from the same thread, can be concatenated and a regression model of change can be applied for a whole thread.

FoCUS (Forum Crawler Under Supervision) was presented in (Jiang et al., 2012). Their goal was to build a forum crawler for visiting only relevant content with minimal overhead. FoCUS design was based on the observation, that, despite being generated by various software engines, having different styles and structures, web forums have similar, so called, implicit navigation paths, represented by specific URLs, that connect entry pages and thread pages. Authors reduced the forum crawling problem, to the problem of recognition of URL type, and enabled the crawler to learn regular expression patterns of the implicit navigation paths. It was showed that, with the use of robust page classifier, even small training set can be enough for a large scale forum crawling.

2.5 Motivation

As evident from the above literature review, there is abundance of methods for scheduling web pages for retrieval. However we would like to validate the basic assumptions and see what are the actual patterns of web page changes on web forums, which are our target. The experiments are performed to further refine the scheduling methods based on the web page changes or propose some hybrid approach, which would take into account a larger number of factors, leading to even better crawler performance.

3 EXPERIMENT SETUP

3.1 What is a Change?

To analyze the data we had to define the notion of page change. The data that we gathered at certain time points contained three types of information: the page was not available under a given URL (marked N/A), the page was available and did not change in comparison to the previous fetch of the page from a given URL even by a single byte (marked as 0), or the page was available and comparison with the previous fetch revealed a change (even if only by a single bit) – marked as 1. The latter case includes also a situation, when the page was collected for the first time, or it was not available under a given URL previously, but at certain point in time a non-empty content was fetched from this URL.

Therefore several situations may occur in practice:

- N/A-N/A – page not available previously and in a subsequent fetch, no change
- N/A-1 – page not available previously, but content fetched in a subsequent round
- 1-1 – page changed previously and in a subsequent fetch
- 1-N/A – page was available previously but disappeared in a subsequent fetch
- 1-0 – page was available previously, and did not change
- 0-0 – page was available previously and still did not change
- 0-N/A – page was available and disappeared
- 0-1 – page was available and changed in a subsequent fetch

Out of all possibilities one (N/A-0) could not occur in our data set.

The motivation for this method of counting changes is that we wanted to distinguish between the pages that appeared and did not change later and the pages that appeared and disappeared immediately (or even later in time).

3.2 Data Set Characteristics

For our studies we have chosen 16 different Polish forum sites, which were periodically crawled during the experiment. After each crawl, a new snapshot was created, that consisted of files fetched from a given web site. The files were stored in a raw form for the ease of further analysis and comparison.

We wanted to achieve a reasonably high granularity of fetches, so the web pages from specified sites were retrieved every 2 hours. As the experiment lasted for 23 days (2012-05-05 – 2012-05-27), this gave us a sample of 266 snapshots for each forum site. The total number of unique URLs that were visited is 27958, including web pages that were discovered during the experiment, as well as pages that, after initial retrieval, were found unavailable later on. To determine the existence of change (as described in 3.1) we used data from 649705 successful fetch events that were conducted.

Similarly to the (Cho and Garcia-Molina, 2000), we used an active crawling technique with a page window. Using active crawling, a program robot (crawler) visits selected web pages periodically and each version of a page is stored for further analysis and change detection. Active crawling can be too obtrusive for web sites being monitored, especially if the fetch frequency is high. Nevertheless, we chose this approach because it provides more reliable statistics and can be more precisely customised in terms of frequency of visits and the scope of monitoring.

In our case the selection of web pages to be visited depends on the structure of a given web site. For each web site the crawler starts with a predefined main page, and conducts breadth-first search for available web pages. All pages that can be found not deeper than on the fifth level of the web site, are perceived as a page window for that site. The page window can change during the consecutive crawls – some pages can be created ore moved to an appropriate level, and thus can be added to the page window; other ones can be removed from the window as they become unavailable or as they are moved deeper in the web site structure.

This method of crawling for the depth of five allowed us to finish each crawl before the time point to start the subsequent crawl. Obviously, we could extend the depth of the crawl, but given our infrastructure limitations, it would require extending the time span between subsequent crawls as well. Our initial experiments indicated, that for the depth of 6 or more, the number of pages retrieved in each crawl raises significantly, which was expected.

4 RESULTS

Our studies show some interesting features of web forum sites that were monitored. The 1 shows the number of new and deleted web pages, that were found during the experiment period, aggregated on daily basis. The data indicate a very high change rate of the local collection – there was more than 1000 new pages on average every day, and a similar amount of web pages were found to be no longer available. As the average amount of web pages in the local collection was 3028, that means every day 1/3 of the collection was replaced by new pages.

It can be seen, that the number of both, the new and the deleted web pages, prove to have weekly fluctuation, with the peak in the middle of the week, and decreasing at the weekends.

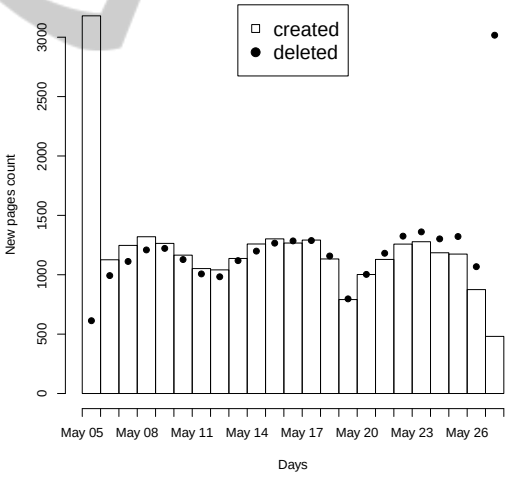

Figure 1: Number of new URLs over the experiment period per day.

The 2 shows the results of our experiment concerning average change frequency of retrieved web pages. As it can be seen, a significant percent of pages change very often — 30.42% of monitored web pages changed every 2 hours or more often on average, 50.41% – at least every 4 hours on average and 87.96% every day or more often on average.

To analyze how quickly the local collection becomes out of date, we measured the time from the

first fetch of each web page to its first change afterwards. A snapshot of web pages that was retrieved from web forums, became obsolete in a short time – after 2 hours, 40.49% of web pages were out of date, after 4 hours – it was 63.6%, and it took only 12 hours for 80.56% of web pages to be outdated.

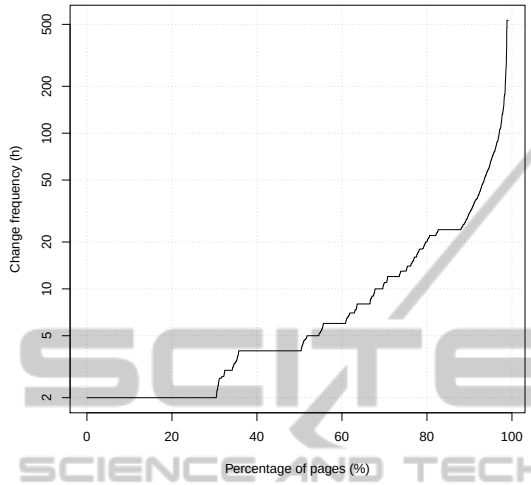

Figure 2: Percent of pages with a given change frequency rate (log scale).

We also calculated the lifespan of web pages in several ranges, as the time span between the first date and the last date when the content was available under a given URL, regardless of its changes and possible disappearances in the meantime. The data presented in 1 indicates that a large number of pages has relatively short lifespan (over 40% disappears within 4 hours). At the same time almost 25% of pages lives longer than one day. Obviously our crawling approach with page window could affect these results to an unknown extent.

Table 1: Lifespan of web pages.

Time	Percentage
0-2 h	27.82%
2-4 h	14.9%
4-8 h	8.16%
8-12 h	8.54%
12-24 h	16.44%
1 day – 1 week	14.16%
> 1 week	9.95%

Finally, we analysed the relation between frequency of changes and possible web page behaviour patterns approximated by the standard deviation of the frequency of changes. We have found that the pages that change frequently (with averages between 1 to 10 hours) present consistent behaviour in terms of their frequency of changes. The higher average time span between subsequent changes, the less consistent behaviour one page presents - standard deviations tend to fan out for higher averages. This might be an evidence for low predictability of web page change patterns and randomness of the change occurrences.

We speculate, that the group of pages that changes often and consistently include the "front" of the web forum, where new information is published, while the other group of pages, that change much less often, consist of pages that are updated, removed or commented. A relatively broad range of most frequently changing pages indicates also, that the new posts pushing back the old ones create artificial change notifications perceived on the web page level. To capture actual changes of content more in depth analysis of the pages is necessary.

5 SUMMARY OF THE EXPERIMENT RESULTS

Over 30 % pages are worth visiting more often than every 2 hours, and over 50 % of pages are worth visiting more often than every 4 hours. The open question is how much of these 30 % are changed due to scripts generating the page, without any actual new content. However, regardless of the answer to this question, which would require another experiment with much higher granularity of fetches, it seems that for instant monitoring of forums it is essential to identify frequently changing pages, in order to capture new content as soon as it appears, and be able to increase the freshness of the collection.

6 IMPLICATIONS FOR CRAWLER DESIGN

Our main motivation to conduct the experiments was to check, whether existing incremental crawling approaches are justified and sufficient for dealing with web forums. The data that we obtained indicate, that if large scale forum crawling is to be performed, high frequency of changes of the web pages has to be taken into account. Our goal for crawling is to be able to capture new information as early as possible. Since large majority of web pages changes within 24 hours it is necessary for the crawler to revisit URLs much more often. Given the usual situation of limited network bandwidth and the necessary politeness policy to avoid overburdening of target web servers it is essential to guide the crawl towards the resources that change more often and are of particular relevance.

However, randomness of the web page change process causes, that basing web page revisit on an estimate of web page change frequency might not be sufficient. Other factors could be taken into account, such as a day of a week or a forum structure. Combining information about the frequency of changes together with information about the structure of a web site as a graph could indicate places in the graph, where new information is actually published. At the same time it would allow to optimize the revisit policy to be able to visit most relevant URLs on a less-than-hour basis.

ACKNOWLEDGEMENTS

The work published in this article was supported by the project titled: "Ego – Virtual Identity", financed by the Polish National Centre of Research and Development (NCBiR), contract no. NR11-0037-10.

REFERENCES

Adar, E., Teevan, J., and Dumais, S. T. (2009a). Resonance on the web: web dynamics and revisitation patterns. In *Proceedings of the 27th international conference on Human factors in computing systems*, CHI '09, pages 1381–1390, New York, NY, USA. ACM.

Adar, E., Teevan, J., Dumais, S. T., and Elsas, J. L. (2009b). The web changes everything: understanding the dynamics of web content. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 282–291, New York, NY, USA. ACM.

Baeza-Yates, R. and Castillo, C. (2007). Crawling the infinite web. *J. Web Eng.*, 6(1):49–72.

Baeza-Yates, R., Castillo, C., Marin, M., and Rodriguez, A. (2005). Crawling a country: better strategies than breadth-first for web page ordering. In *Special interest tracks and posters of the 14th WWW conference*, WWW '05, pages 864–872, New York. ACM.

Ben Saad, M. and Gançarski, S. (2011). Archiving the web using page changes patterns: a case study. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 113–122, New York, NY, USA. ACM.

Buttler, D., Rocco, D., and Liu, L. (2004). Efficient web change monitoring with page digest. In *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, WWW Alt. '04, pages 476–477, New York, NY, USA. ACM.

Cai, R., Yang, J.-M., Lai, W., Wang, Y., and Zhang, L. (2008). irobot: an intelligent crawler for web forums. In *Proceedings of the 17th WWW conference*, WWW '08, pages 447–456, New York, NY, USA. ACM.

Cho, J. and Garcia-Molina, H. (2000). The evolution of the web and implications for an incremental crawler. In *Proceedings of the 26th International Conference on Very Large Data Bases*, VLDB '00, pages 200–209, San Francisco. Morgan Kaufmann Publishers Inc.

Cho, J. and Garcia-Molina, H. (2003). Estimating frequency of change. *ACM Trans. Internet Technol.*, 3(3):256–290.

Douglis, F. and Ball, T. (1996). Tracking and viewing changes on the web. In *USENIX Technical Conference*. AT&T Bell Laboratories.

Douglis, F., Ball, T., Chen, Y.-F., and Koutsofios, E. (1998). The AT&T Internet Difference Engine: Tracking and viewing changes on the web. *World Wide Web*, 1:27–44.

Hirschberg, D. S. (1977). Algorithms for the longest common subsequence problem. *J. ACM*, 24(4):664–675.

Jiang, J., Yu, N., and Lin, C.-Y. (2012). Focus: learning to crawl web forums. In *Proceedings of the 21st WWW conference*, WWW '12 Companion, pages 33–42, New York. ACM.

Kwon, S., Lee, S., and Kim, S. (2006). Effective criteria for web page changes. In Zhou, X., Li, J., Shen, H., Kitsuregawa, M., and Zhang, Y., editors, *Frontiers of WWW Research and Development - APWeb 2006*, volume 3841 of *Lecture Notes in Computer Science*, pages 837–842. Springer Berlin / Heidelberg.

Law, M. T., Thome, N., Gançarski, S., and Cord, M. (2012). Structural and visual comparisons for web page archiving. In *Proceedings of the 2012 ACM symposium on Document engineering*, DocEng '12, pages 117–120, New York, NY, USA. ACM.

Liu, M., Cai, R., Zhang, M., and Zhang, L. (2011). User browsing behavior-driven web crawling. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 87–92, New York, NY, USA. ACM.

Rocco, D., Buttler, D., and Liu, L. (2003). Page digest for large-scale web services. In *IEEE International Conference on E-Commerce*, pages 381 – 390.

Saad, M. B. and Gançarski, S. (2010). Using visual pages analysis for optimizing web archiving. In *Proceedings of the 2010 EDBT/ICDT Workshops*, EDBT '10, pages 43:1–43:7, New York, NY, USA. ACM.

Toyoda, M. and Kitsuregawa, M. (2006). What's really new on the web?: identifying new pages from a series of unstable web snapshots. In *Proceedings of the 15th WWW conference*, WWW '06, pages 233–241, New York, NY, USA. ACM.

Yang, J.-M., Cai, R., Wang, C., Huang, H., Zhang, L., and Ma, W.-Y. (2009). Incorporating site-level knowledge for incremental crawling of web forums: a list-wise strategy. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 1375–1384, New York, NY, USA. ACM.