The WT10G dataset and the evolution of the Web

Wei-Tsen Milly Chiang
University of Wollongong
Wollongong, Australia
wmc01@uow.edu.au

Markus Hagenbuchner
University of Wollongong
Wollongong, Australia
markus@uow.edu.au

Ah Chung Tsoi
Australian Research Council
Canberra, Australia
ahchung.tsoi@arc.gov.au

ABSTRACT

The purpose of this paper is threefold. First, we study the evolution of the web based on data available from an earlier snapshot of the web and compare the results with those predicted in [2]. Secondly, we establish whether the WT10G dataset, a popular benchmark for the development and evaluation of internet based applications is appropriate for the tasks. Finally, is there a need for a collection of a new dataset for such purposes. The findings are that the appropriateness of using the popular WT10G dataset in recent Internet-based experiments is questionable and that there is a need for a new collection of dataset for development and evaluation purposes of algorithms related to Internet search engine developments.

Categories and Subject Descriptors: H.5.4 Information Interfaces and Presentation: Hypertext/Hypermedia. H.4.m Information Systems: Miscellaneous.

General Terms: Experimentation, Verification.

Keywords: Web evolution, rate of change, standard datasets.

1. INTRODUCTION

A web repository is a closed set of web pages retrieved from the Internet over a short period of time to ensure that the varying nature of the web does not introduce significant effect on the collected data. There are numerous research areas which utilize a web repository to evaluate new proposed algorithms and approaches. Due to the ever changing nature of the world wide web, new models are often developed on a static snapshot of the web. The use of static data allows to study and evaluate the proposed algorithms and hence, allows reproducible experiments to be conducted within a controlled environment. A common dataset used by researchers for such purposes is the WT10G (Web Track 10Gigabytes) [1] distributed by CSIRO in Australia. It has been an ideal testbed for many web-based experiments. The WT10G collection is a subset of a larger collection known as WT100G or VLC2, developed from a crawl of the web content in 1997. The WT10G collection was constructed in a way which retains the properties of the 1997 web content, therefore allowing conclusions drawn from experiments on such a collection to be applicable on the world wide web. The properties retained in WT10G include: the web link graph structure, server size distribution, inclusion of inter-domain links and inclusion of web pages on various subjects [1]. The popularity of the dataset is reflected in a preliminary investigation on published literatures which identified 43 papers which employed the WT10G or WT100G data collection to evaluate their approaches. We also found that the number of publications based on the WT10G or WT100G web repository increases over time from 1 in 2000 to 17 in 2003 and 2004 respectively.

WWW 2005, May 10–14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

	WT10G	Expected
Number of domains	11672	
Domains accessible	5984 (51.27%)	91 (0.781%)
Number of pages	1692096	
Pages accessible	5884 (0.35%)	17 (0.001%)

Table 1: Accessibility of pages and domains in the WT10G collection. Valid as of September 2004.

The Web has changed significantly since 1997 and hence a question arises to whether the WT10G dataset still is a valid representation of the web and its properties. A study conducted in 2004 on the evolution of the web [2] found some significant dynamics. [2] focused on the investigation on the accessibility of web pages and the variation of web contents over a period of time. We will evaluate the characteristics of the WT10G against the 2004 web with respect to the criteria, to see if the computed average changes match those predicted in [2].

2. WEB PAGE ACCESSIBILITY RATE

A recent study [2] shows that approximately 50% of existing domains, and 80% of web pages will no longer be accessible after one year. We compared these figures with those of the WT10G dataset. For this purpose, the domains and web pages contained in WT10G were assessed for their accessibility in the current Web by establishing a connection with the corresponding URL. The accessibility was determined by analyzing the response codes received from the connections. Some problems were encountered during assessment: page redirection, automatic generated domain re-selling pages. Affected domains or pages were not considered valid.

As Table 1 shows, of the 11672 domains in WT10G, 5984 (51%) were still valid in September 2004. The value in the expected accessibility column is derived from the study in [2], which observed that approximately 50% of domains are not accessible after one year. Based on the assumption of a linear change in web contents, as observed in the experiment by the Online Computer Library Center for the period of 1998 to 2002 [2], a compound rate method is used to project that a collection from 1997 should have approximately 0.781% of the existing domains accessible in 2004.

Of the 1692096 pages in WT10G only 5884 (0.35%) pages were valid in September 2004. The number of pages accessible is even lower than the number of accessible domains. This peculiar finding could be attributed to the large amount of change that has occurred in the 7 years since, where the same domain may still exist, but the pages may have different naming schemes or extensions. The expected accessibility identified in [2], where web pages were observed to become inaccessible at a rate of 80% per year, imply an

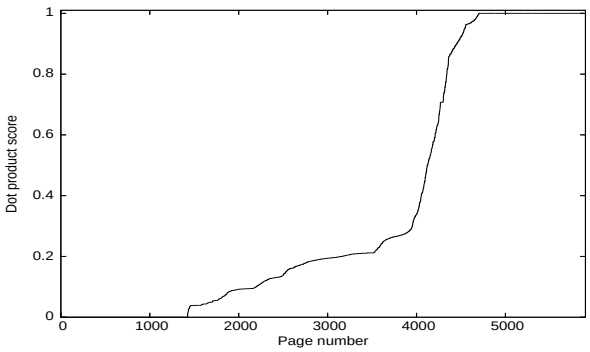

Figure 1: The vocabulary vector dot product score

accessible rate of 0.001% after 7 years. Again, our empirical assessment does not agree with the rate predicted in [2].

These findings show that contrary to prediction, the dynamics of the World Wide Web has not accelerated as fast as predicted by [2]. However, given that dynamically created, user customized pages become commonplace, it can be expected that the rate of changes in the Web will grow further.

3. VARIATION OF WEB CONTENT

The popularity of the Internet has encouraged a wide variety of web content; therefore, there is a need to examine the amount of content change that occurred in web pages in order to determine whether the WT10G collection continue to include an appropriate variety of subject areas. To evaluate the amount of subject change in web page content, an empirical assessment was conducted by analyzing the 5884 web pages from the WT10G collection which were still accessible in September 2004. This content evaluation assessment examined the change in content subject by extracting a vector of vocabularies from both the 1997 and 2004 versions of the web page residing in the same web address. Although 5884 were qualified, after the vocabulary extraction process, 141 pages with an empty vocabulary vector were removed. Then the dot product of the respective vocabulary vectors of the 1997 and 2004 versions is taken to yield a score that ranges from 0 to 1, where 0 indicates a total change of subject and 1 indicates no change.

From Figure 1, it can be observed that the web pages fall into 3 areas. From the 5743 valid pages, 1285 (22.38%) have a score of 0, which indicates that these pages have a totally different content in 2004. On the other hand, there are 798 (13.90%) pages with a score of 1, which remains the same even after the 7 year period. To further understand the changes that occurred in these webpages, the number of pages with a significant change (a dot product score of < 0.3) of content is identified. It is observed that more than two-thirds of pages have had a significant change in content, implying a change in the subject area of those web pages.

We are aware that in some cases, the nature of a web page is to be frequently updated with a variety of content, such as those for news publishing purposes; therefore, we have conducted a follow-up investigation that looks at the number of unique words in a web page. The investigation revealed that more than a half of the pages have had a dramatic growth or reduction in page content, where the difference in the number of unique words is more than 50. Approximately 59% of those pages had an increase of unique words, and 34% experienced a decrease. This suggests that a large proportion of web pages are updated with a content that provides more information. In addition, there were 1134 pages with a dot product score of 0.3 to 1, and which remained focused in a similar subject area. 10% of these pages had a difference of more than 50 in the number of unique words, indicating a significant addition of information on existing content or broadening of subject area in the web page content.

The observation that a large proportion of websites have undergone a significant change in content could be a result of the introduction of new subject areas over the years and the increasing popularity of website maintenance tools. These new characteristics of the web are not reflected in the WT10G collection, therefore confirming the need for a new yet comparable collection.

4. A NEW WEB REPOSITORY

We are not aware of any other publicly available web repository of reasonable size other than the WT10G or WT100G [1]. We have demonstrated that the WT10G dataset may be out-of-date, suggesting that it is time to establish a new dataset. Our research group is in the process of crawling a portion of the world wide web with the aim to establish an alternative to the WT10G dataset. The dataset will be made available via the web site www.artificial-neural.net.

The aim is to construct a web repository that is comparable in effectiveness to that of the WT10G collection while addressing the associated issues as those identified in previous sections. To achieve this, a procedure similar to that used for the construction of the WT10G collection will be executed. Special care is taken to ensure that a balanced proportion of pages are crawled with coverage of a wide variety of subjects including new or emerging topics in the current world wide web. The format of the new dataset will be compatible with that of the WT10G dataset. Such a dataset will be specifically suitable in the evaluation of internet search engine algorithms based on link and/or content analysis techniques.

5. CONCLUSIONS

This paper has provided evidence to raise doubt about the WT10G collection. The paper also found that some assumptions of existing models about the dynamic of the web are inaccurate since the rate of changes in the web do not appear to be linear but rather a slower accelerated rate of changes is observed. It can be assumed that this slower acceleration will continue in the future.

The broader research community benefits from the availability of standard datasets. Since the WT10G repository is questionable as a suitable database for experimental evaluations, a lack of a suitable publicly available benchmark web repository is identified. A current effort is to make a new dataset available which is specifically suited to the evaluation of internet search engine capabilities using link and/or content analysis of web pages.

Future tasks include the development of a more accurate prediction model on the evolution of the world wide web. This can be of great values when designing datasets which consider the dynamics of the web.

6. ACKNOWLEDGMENTS

The first and second author acknowledge financial support from the Australian Research Council in the form of a Discovery Project grant.

7. REFERENCES

[1] D. Hawking. *Web Research Collection*. http://es.csiro.au/TRECWeb/, June 2004.

[2] A. Ntoulas, J. Cho, and C. Olston. What's new on the web? the evolution of the Web from a search engine perspective. In *Proceedings of the 13th WWW Conference*, pages 1–11, 2004.

[3] R. Roach. Survey reveals 10 biggest trends in Internet use. *Black Issues in High Education*, page 42, Oct 2004.

[1]There are a number of web repositories collected by individual groups, but these are not normally publicly available.