Statistical Study of the Thai Web Structure

Piyapong Seibudta, Wiwat Nimsanitnan, Bundit Manaskasemsak, Arnon Rungsawang

MIKE *Laboratory*, Department of Computer Engineering, Faculty of Engineering,

Kasetsart University, Bangkok 10900, Thailand.

piya\_tt@hotmail.com, wiwat@octknight.com, {un,arnon}@mikelab.net

Abstract

Studying of a national web structure can provide an insight into society specific to a country. In this paper, we propose a study of the Thai web structure. A sample set of Thai web pages have been crawled and analyzed. We first extract statistical and structural properties from the Web data. We then examine its link statistics, and found similar characteristics to those of other country web structures. We further analyze and report its geo-location, top-level domain (TLD), and content statistics that are related to the Thai web structure.

**Keywords:** web structure, web statistics, web content

# 1. Introduction

Access to the Internet concerns everyday life in our education, commerce, business, recreation, etc. Many web sites have been unceasingly created. Owing to the Netcraft survey [7], over nine hundred million web sites were recorded in March, 2014. This number has been still increasing; the web content is also dynamically changing and updating as well. Studying the evolution of the web structure by its statistic is technically challenging and practically useful in development of efficient algorithms for many web applications, such as web crawling, web search and navigation etc.

This paper presents the statistical study of a sample of the Thai web structure. We define the Thai web as a set of web pages related to Thailand. While we first confirm link structure properties already observed in other country web graph studies [1, 5-6] (i.e., power-law distribution, bow-tie structure), this statistical study reveal some characteristics that specific to Thai web.

# 2. Thai Web Dataset

Our Thai web dataset is defined as web pages of all web sites that are either registered under.th domain, or hosted at the IP addresses associated with Thailand. We also include those web pages in .com, .net, and etc. that are obtained from a Thai language-specific web site crawling result of Tadapak et al. [8] We launch the breadth-first search crawler to construct a sample of the Thai web dataset, using the seed sets from Tadapak et al.’s recommendation and consider only web pages that have found at least 10% of Thai words. The resulting web dataset contains 61,821 sites, 3,985,981 pages.

# 3. Thai Web Structure

**3.1 Link Statistics**

Host graph is defined as a directed graph where each node represents a web site (or host), and each directed edge represents hyperlinks from web pages on the source web site to those on the target one. Following the host-graph analysis in [2], we found that our web dataset shows characteristics that are very similar to those of other country web studies [b-d]. The in-degree and out-degree distributions in the log-log plots in Fig. 1 can be approximately by a straight line, a signature of the power-law distribution. In addition, the result from connected component analysis in Fig. 2 also shows a bow-tie structure [3]. Thus, the statistical study from this sample web dataset can represent a big picture of the whole current Thai web structure.

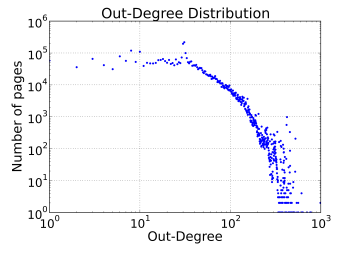
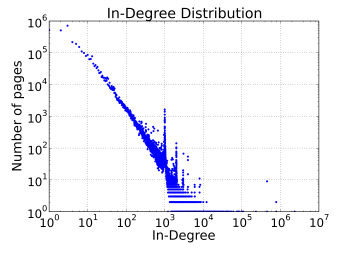


Figure 1: In-degree and out-degree distribution.

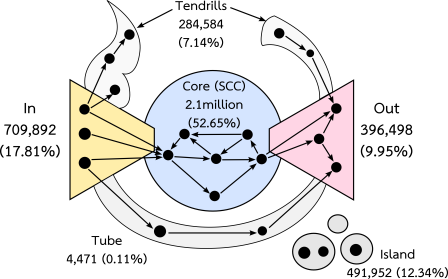


Figure 2: Bow-tie structure of the Thai web.

When we examine the PageRank scores of web pages in the dataset (cf. Fig. 3), we found less number of web pages has high PageRank scores, while most of them have lower scores. This result confirms the emergence of the power-law distribution of the PageRank scores, as well.

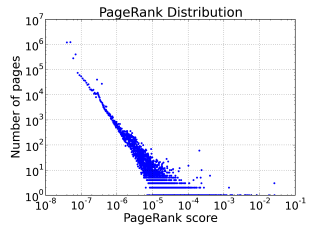


Figure 3: Distribution of the PageRank scores.

**3.2 Geo-location and TLD Statistics**

When we check the geographical location of an IP address of each web server, we found 93.14% of web sites are in Thailand, while 6.86% of the rest situate in foreign countries. Fig. 4 concludes only the percentage ratio of foreign countries that the Thai web servers are located. We found most web servers have registered their geo-location in the US.

Figure 4:Web structure by geo-locations.

When we analyze the whole web structure by TLD (Top Level Domain), we actually found larger number of web sites in .com more than .th domain. We also found many web sites in .net and .org, respectively. Fig. 5 concluded this finding.

Figure 5: Web structure by TLDs.

**3.3 Content Statistics**

Following web categories recommended by Dmoz [4], we classify the web sites and obtain categorical result shown in Fig. 6. We found many Thai web sites in computer, social, and business categories.

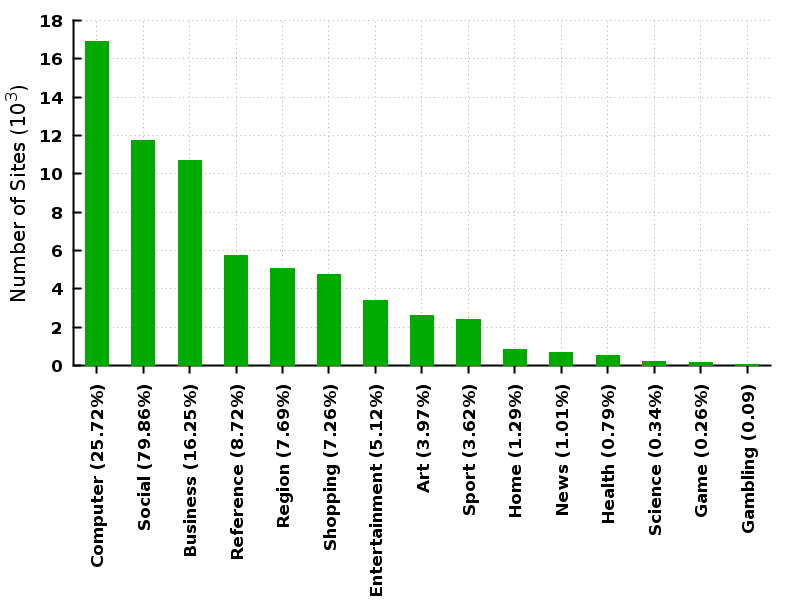


Figure 6: Web structure by content categories.

# 4. Conclusion

In this paper, we propose a study of web structure induced from a sample set of Thai web pages. We analyze connectivity characteristics, and found that this web dataset can represent the perspective of the whole Thai web structure. We also examine and report the geo-location and TLD, as well as the content statistics of the sample dataset. These statistics can help in the design of more efficient web crawling, web searching and navigation algorithms, especially to the Thai web structure. In the future, we plan to study the evolution of the Thai web using future crawled archival data, and analyze the different web snapshots.

# References

[1] R. Baeza-Yates, C. Castillo, and V. Lopez, “Characte-ristics of the Web of Spain”, *Cybermetrics*, 9(1), 2005.

[2] K. Bharat, B. Chang, M. Henzinger, and M. Ruhl, “Who links to whom: Mining Linkage between Web Sites”, *Proc. of Inter. Conf. on Data Mining*, 2001.

[3] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, and A. Tomkins, “Graph Structure in the Web”, *Computer Networks*, vol.33, 2000.

[4] Dmoz, http://www.dmoz.org/World/Thai

[5] I. Han, S. H. Lee, and S. Lee, “Graph Structure of the Korea Web”, *Proc. of the 12th Inter. Conf. on Database Systems for Advanced Applications*, 2007.

[6] G. Lui, Y. Yu, J. Han, and G. Xue, “China Web Graph Measurements and Evolution”, *Proc. of the 7th Asia Pacific Web Conference*, 2005.

[7] Netcraft, http://news.netcraft.com

[8] P. Tadapak, T. Suebchua, and A. Rungsawang, “A Machine Learning Based Language Specific Web Site Crawler”, *Proc. of 13th Network based Information System*, 2010.