Facts about the Thai Web

Surasak Sanguanpong
Department of Computer Engineering
Kasetsart University
Bangkok, Thailand
E-Mail: nguan@ku.ac.th

Suthiphol Warangrit and Kasom Koth-Arsa
Department of Computer Engineering
Kasetsart University
Bangkok, Thailand
E-Mail: g4165234, g4265077@ku.ac.th

Abstract: This paper presents the latest status of Thai web servers. Quantitative measurements are based on database crawling on July 2000. Our experiment shows that the Heaps' and Zipf's laws apply strongly to documents on the Thai web. A visualization tool is developed to show servers connectivity.

Keywords: spider, robot, Thai web, Heaps' law, Zipf's law

1. Introduction

The ultimate growth of the Internet and WWW is well-known fact. In Thailand, the major growth is contributed by academic and commercial sectors. From our knowledge, the only summary numbers related to the web servers in Thailand is presented in [3]. More complete and up-to-date statistics about the Thai web can be found in [4].

In this paper, we summarize the latest status of Thai web based on databases crawling on July 2000. The collection was run using *NontriSpider* [4] built on a 500 MHz Pentium III Xeon PC, 512 MB of RAM, 18.2 GB of local disk, and a 155 Mbps ATM connection to the Internet. Running on this platform, *NontriSpider* collected nearly 1.1 million HTML documents in 8 hours.

We apply the Heaps' and Zipf's law to model characteristics of documents. Furthermore, we have developed a tool for visualization of connectivity between these servers.

2. Statistics Related to the Thai Web

We found 1,086,846 HTML documents on 6,763 web servers, and 2,575 third level domain names. The total size of HTML documents is 12.2 GB. All servers are categorized into eight major domains as shown in Table 1.

3. Document Characteristics Modeling

We assume some laws widely accepted in Information Retrieval, which are shown valid in our experiments. The first one is generalized Zipf's law [1], which attempts to capture the relationship between a word's popularity in terms of rank and its frequency of use. It states that if one ranks the popularity r of words used in a given text by their frequency of use P then $P \sim 1/r^{\theta}$ with the exponent θ is often close to unity. Figure 1 (a) illustrates the distribution of frequencies considering that the word are arranged in decreasing order of their frequencies. In our case, we get $\theta = 0.85$.

The next characteristic is the growth of vocabulary as a function of the text size. Heaps' Law [2] is used to predict the growth of the vocabulary size. The law states that the vocabulary of a text size n words is of size $V = O(n^{\beta})$, for $0<\beta<1$. Figure 1 (b) illustrates how the vocabulary size varies with the page size. In our experiment, we get $\beta = 0.5$.

Table 1 : Number Related to the Thai Web Size

Domain	#Web Servers	#HTML documents
ac.th	2,166(32.03%)	500,260 (46.03%)
co.th	3,260 (48.20%)	267,759 (24.63%)
go.th	433 (6.40%)	144,169 (13.26%)
in.th	215 (3.17%)	6,679 (0.61%)
mi.th	20 (0.30%)	6,379 (0.59%)
net.th	186 (2.50%)	63,667 (5.86%)
or.th	482 (7.13%)	97,776 (9.00%)

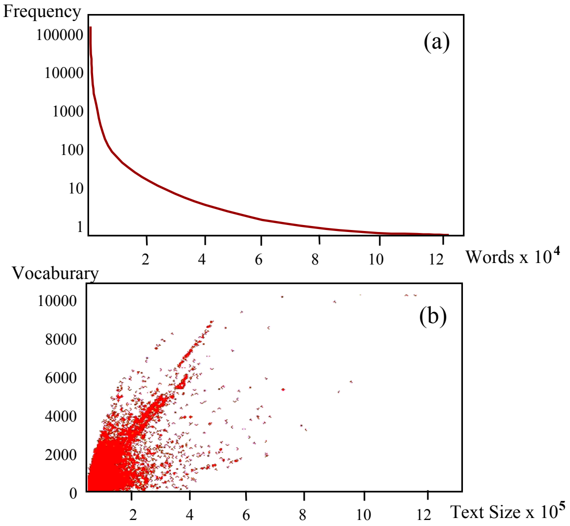

Figure 1: Distribution of sorted words frequencies (a) and size of the vocabulary (b)

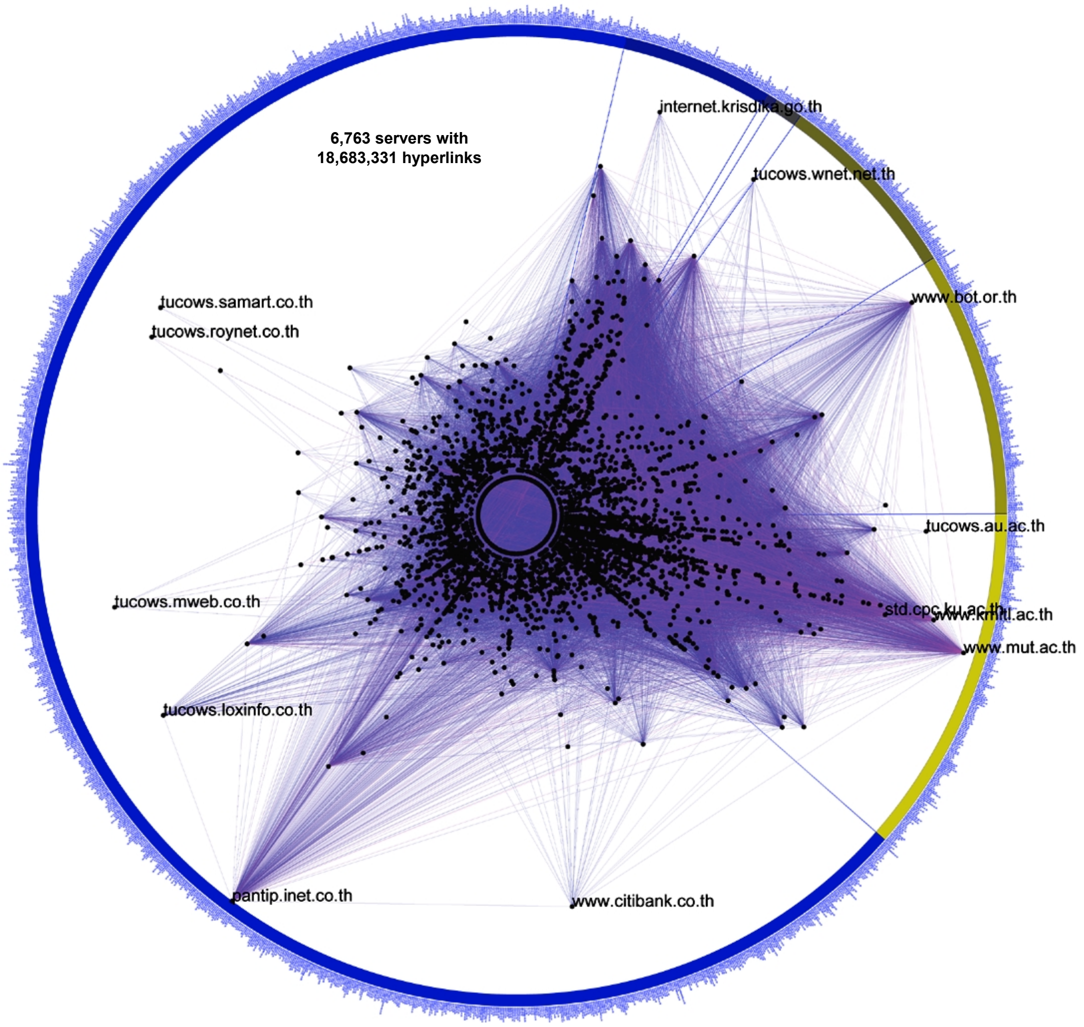

Figure 2: Visualization of Servers Connectivity

4. Visualization of Connectivity

We develop a hyperlinks visualization tool called *Hyvis*. A matrix of hyperlinks is first built, then the lines were drawn between servers that have links to each other. Figure 2 shows a two-dimensional visualization of *Hyvis* data depicting a snapshot of servers connectivity.

The graph reflects 6,763 servers with 18,663,331 links. The space was divided into eight sections corresponds to eight domains. Each section occupies the space proportional to the number of domains it has. All servers are plotted along the radius. Their positions are at the radius R = (*number of pages*/*the maximum number of pages*)$^{0.25}$.

5. Conclusions

This paper presents a summary concerning the Thai web, and attempts to provides some quantitative answers to the question about the Thai web size. It uses the numbers to drive a visualization of servers connectivity. Our experiment shows that the Heaps' and Zipf's laws apply strongly to HTML documents in the Thai web.

6. Acknowledgment

The authors thank Dr. Punpiti Piamsa-nga for his comments on the manuscript. This work is funded by the KURDI research grants SRU-2.43

7. References

[1] Gonnet, G., and Baeza-Yates, R. *Handbook of Algorithms and Data Structures* 2nd edition, Addison-Wesley, England 1991.

[2] Heaps, J. *Information Retrieval–Computation and Theoretical Aspects,* Academic Press, 1978.

[3] Roehrl, A., Frey, M., and Roehrl, R. *World Wide Web Robot for Extreme Datamining with Swiss-Tx Supercomputers,* Interim Report IIASA IR-99-20, 1999.

[4] Sanguanpong, S., Piamsa-nga, P., Poovarawan, Y., and Warangrit, S. *Measuring Thai Web using NontriSpider*, Proceeding of the International Forum cum Conference on Information Technology and Communication :123-132, June 2000.