Measuring and Analysis of the Thai World Wide Web

Surasak Sanguanpong, Punpiti Piamsa-nga,
Somnuk Keretho, Yuen Poovarawan, and Suthiphol Warangrit

Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok 10900, Thailand
E-Mail: {nguan, pp, sk, yuen, g4165234}@ku.ac.th

Abstract

This paper presents quantitative measurements and analyses of various issues related to web servers and web documents in Thailand. We develop a web spider to automatically collect web documents. Using our web spider, we found that there are over 700,000 documents and over 8000 web servers on Thai web (information located on the servers registered under `.th` domain.). Analysis in this paper was based on database crawling on March 20, 2000.

1 Introduction

This paper presents quantitative measurements and analysis concern of web servers in Thailand. We develop a web spider, called *NontriSpider* to automatically collect web documents *NontriSpider* is a multi-process web spider that is developed as a part of *NontriSearch* search engine [Sanguanpong98]. Analysis in this paper was based on database crawling on March 20, 2000. We present several characteristics of Thai web space and investigate certain properties of documents, such as page size distribution, top ten biggest domains, number and type of embedded images, inbound and outbound link characteristics, language usage, etc. Full details of facts and figures are available at `http://anreg.cpe.ku.ac.th/links/thaiwebstat/index.html`

In the next section, we briefly describe the design of our *NontriSpider*. The following section presents the analyses of Thai web. Finally, we conclude the paper and discuss on future work.

2 Collection Tool and Methodology

Web *spider* or *crawler* or *robot* is one of the key components of search engines. The main task of a spider is to automatically gather HTML documents or web pages from public web servers. A number of spiders will be sent to read and analyze web pages in order to build an index for search engine services. Besides search engine applications, collection of web pages is also very useful for other applications, such as resource discovery, data mining, statistics, etc. Research and development of web spider has its milestone less than a decade [Eichmann94, Pinkerton94]. Currently, there are well over two hundred of spiders available on the web [Koster93].

NontriSpider is a web spider, which utilized multiple of web collectors. A web collector is a software process to collect information from a targeted web server. The design concept of multiple-process of spider could be found in [daSilva98, Roehrl99]. *NontriSpider* is a C/C++ application currently running on a single 500-MHz Pentium III Xeon/Linux PC, which has 512-MB memory, 2 units of 9.1GB of Ultra SCSI-2 hard disks, and a 100-Mbps Ethernet connection to the router. The router is connected to THAISARN network [Koanatrakool94] with 2-Mbps bandwidth. All HTTP traffics from our network to hosts in Thailand are forced to go through cache servers. The cache servers utilize 3 nodes Beowulf clusters. Each node is a Pentium III 500-MHz machine with 1-GB of memory and three of 9.1-GB Ultra SCSI-2 disks running Squid.

We measure the performance of *NontriSpider* with the number of collectors from 10 to 300. We ran 20 times of experiments with the same set of URLs in different periods of day. The best five and the worst five results are discarded. Other remaining results are used to compute average data measurement. Furthermore, to get actual performance of the collection without caching environments, we perform another test by running spiders through a 155 Mbps connection to the university network consortium(UNINET), which has shared 8 Mbps bandwidth to the Internet.

The result in Figure 1 shows that collection rates via cache servers grow somewhat linearly and begin to saturate with the collection rate around 50 documents per second. This is because the cache servers could not handle the rapid requests very well. On the other hand, when the spiders get data directly without cache servers, the collection rate is up to 140 documents per second.

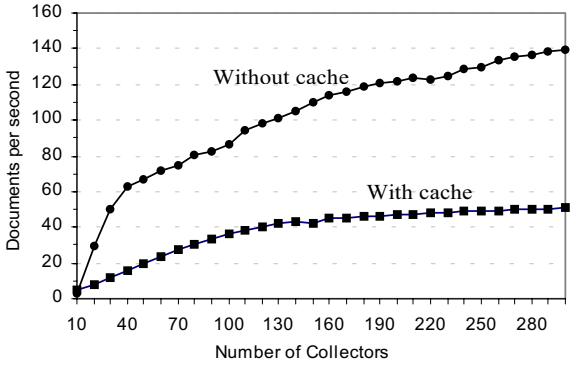

Figure 1: Collection rate of NontriSpider

3 Statistical Analyses of Thai Web

From our knowledge, the information about web servers in Thailand is mentioned in [Roehrl99]. Unfortunately, only summary results were presented without any important statistics. Furthermore, data in [Roehrl99] was obsolete regarding the web explosive growth during the past years. In this paper, we collected web pages and used them to illustrate the current status of Thailand World Wide Web. In the next sub section, we firstly clarify the difference between web servers and web machines. Then, the analysis of experimental results are discussed in eight issues, which are as follows:

1) Size of Thai web
2) Document distribution
3) Number of servers and number of documents in 10 largest domains
4) Page size distribution
5) Number of image counts
6) Behavior of outgoing URLs in documents
7) File type distribution
8) Languages used in documents

3.1 Web Servers and Physical Machines

A web server in this paper is a program that provides HTTP services. It is counted by a unique URL address, not a physical machine. With virtual host technique, one machine may support multiple web servers simultaneously. To precisely get the number of machines that are used as web servers, we count IP addresses of machines rather than their hostnames.

3.2 Size of Thai Web

Five metrics are used to measure the size of Thai web under "`.th`" domain. The metrics are 1) number of HTML documents; 2) size of HTML pages; 3) number of web server addresses; 4) number of web server machines; and 5) number of third-level domains.

Table 1 shows the current measured metrics. We found that there are totally 769,169 HTML documents on 8,520 web servers (average of 90 HTML documents per server). The total size of HTML documents is around 7.6 gigabytes. For physical machines, we found there are 4,420 machines; therefore, in average, each machine supports two web servers. Also, 2830 of third level domain are discovered (compare to 1931 domain reported by [Roehrl99] in 1999). Recent survey shows that there are 71995 Internet hosts in Thailand [IIRC00]. Therefore, 12% of Thai Internet hosts are web servers.

3.3 Document Distribution

Web servers found in Table 1 are categorized into the following sub domains.

- ac.th (academic)
- co.th (commercial)
- go.th (government)
- in.th (individual)
- mi.th (military)
- net.th (network)
- or.th (organization)
- amazingthailand.th[1] (others)

Table 2 shows the number of servers and documents classified in each domain (the domain amazingthailand.th, which has only 1 server and 173 HTML documents, is not shown in this table.) Not surprising that over 80% of web servers and around 72% of all documents are in the academic and commercial domains. Because the Internet in Thailand has begun from academic networks, and the growth is pushed up by the commercial sectors later.

Table 1: Information of web servers under .th domain

Type	Quantity
Number of HTML documents	769,169
Total documents size (GB)	7.6
Number of web servers	8,520
Number of web machines	4,420
Number of third-level domains	2,830

Table 2: Servers and documents classified by domain name

Domain	# servers	# documents
ac.th	2756	382485
co.th	4191	174283
go.th	500	104134
in.th	264	5158
mi.th	24	5538
net.th	218	33010
or.th	566	64388

3.4 Number of Servers and Number of Documents

In this section we use the total number of web servers and total number of HTML documents on each domain to rank Thai web domains, which are illustrated in Table 3 and Table 4, respectively. As illustrated in Table 3, the domain chula.ac.th (Chulalongkorn University) is on top of the list since it has the highest number of web servers. In Table 4, the domain ku.ac.th (Kasetsart University) is the biggest site ranked by the number of HTML documents.

Table 3: Top-Ten domains ranked by number of web servers

Rank	Domain	# servers
1	chula.ac.th	236
2	inet.co.th	210
3	kku.ac.th	163
4	tu.ac.th	149
5	cmu.ac.th	144
6	au.ac.th	132
7	ku.ac.th	125
8	rit.ac.th	124
9	psu.ac.th	115
10	buu.ac.th	107

Table 4: Top-Ten domains ranked by number of documents

Rank	Domain	# documents
1	ku.ac.th	67678
2	inet.co.th	37501
3	kku.ac.th	35333
4	chula.ac.th	28226
5	mut.ac.th	25170
6	wnet.net.th	21650
7	psu.ac.th	21359
8	krisdika.go.th	20958
9	swu.ac.th	20021
10	buu.ac.th	19305

[1] amazingthailand.th is a domain for a special event and is not classified as a standard sub-domain.

Besides information shown in Table 3 and Table 4, the distributions of web servers and HTML documents for all domains are illustrated in Figure 2 and Figure 3, respectively. We found that more than 50% has only one web server in its domain.

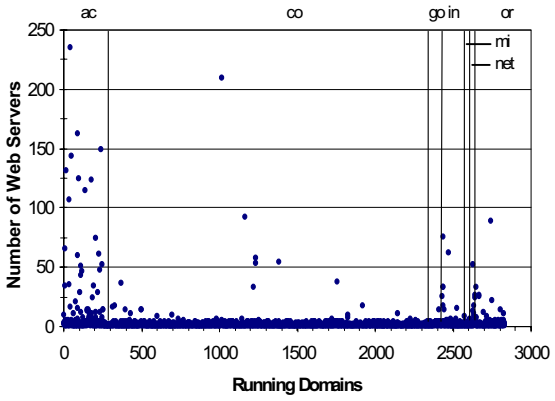

Figure 2: Distribution of web servers

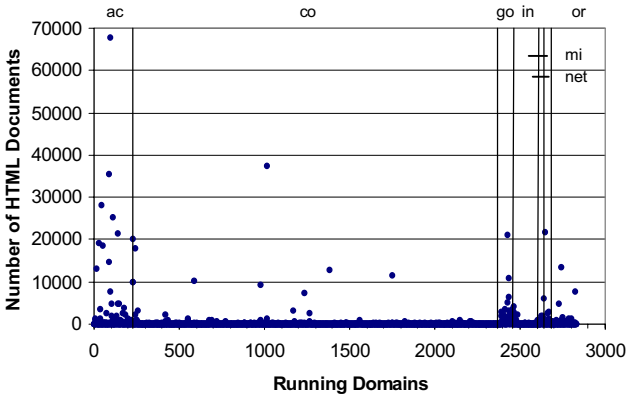

Figure 3: Distribution of HTML documents

3.5 Page Size Distribution

Figure 4 shows the size distribution of the web pages. We can see that majority of pages are small size and it is large enough to contribute significantly to the overall volume observed, i.e., skew the mean page size to 9.8 KB with a very high value of variance (658,640,896 in our experiment).

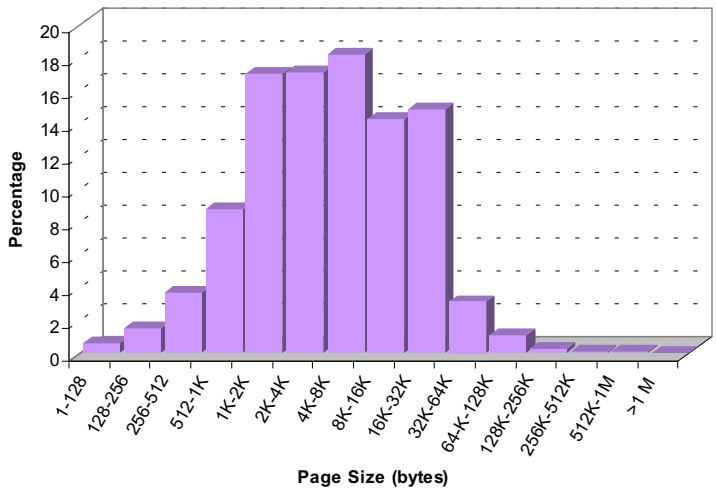

Figure 4: Page size distribution

3.6 Number of Image Counts

Figure 5 shows the distribution of embedded image counts. This figure shows only pages, which contain 0 to 15 images. We found that over 50% of pages contain 1 to 10 images and 24.67% of pages contain no images. Average number of images per page is 10.

3.7 Characteristics of URL Links

Overall characteristics of hyperlinks in Thai web are interesting. Figure 6 shows pages, which has HTML links no more than 30 links. Around 55% of all pages contain 1-10 outgoing links. Around 17% of web pages have no outgoing links. Average number of HTML links in a page is 13 links per page.

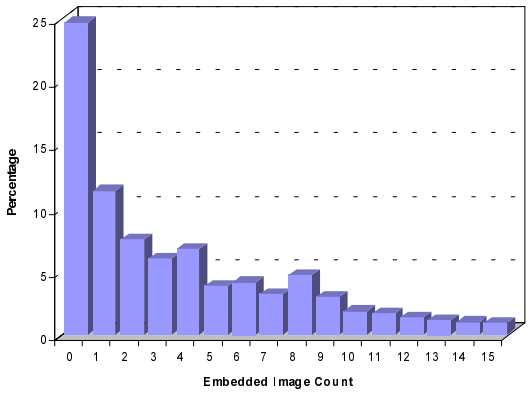

Figure 5: Distribution of image counts

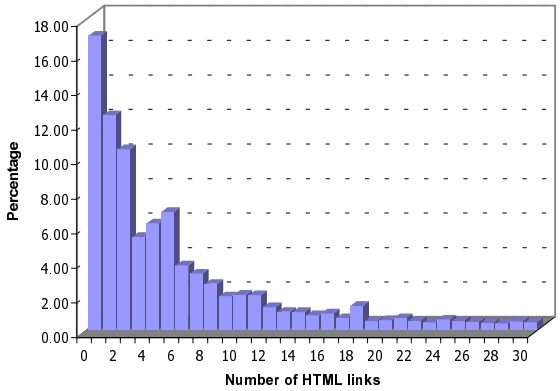

Figure 6: Distribution of outgoing URL

Figure 7 show the analysis of outgoing links, which are classified into 4 categories:

1) Inbound links point into the same web server
2) Inbound links point to different web servers
3) Outbound links point to .th servers
4) Outbound links point to non-Thai domains.

We found the following interesting links characteristics :

- For any web pages, which have less than 40 links/page, 80% of their links falls into the first category.
- For any web pages, which have more than 40 links/page, fewer than 60% of their links are inbound links and the rest are outbound links.
- By manually random investigation, we also found that most of these web pages contain links to software download from foreign sites.

3.8 File Type Distribution

In this section, we investigated the file type distribution. Classification has been done by the standard suffix used in file names e.g. .html, .htm, .jpg, .gif, etc. File names without suffix are classified as unknown. We found that 70% of all file types are HTML documents. The distribution of all file types is shown in Figure 8.

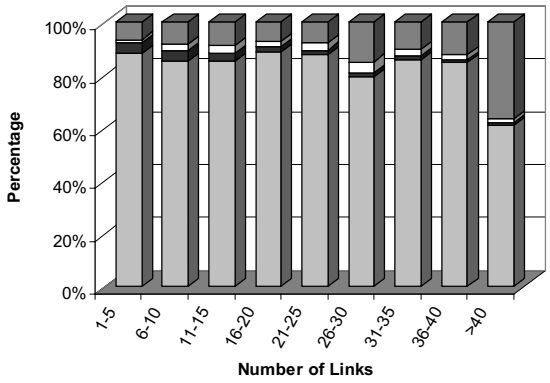

Figure 7 Distribution of outgoing links

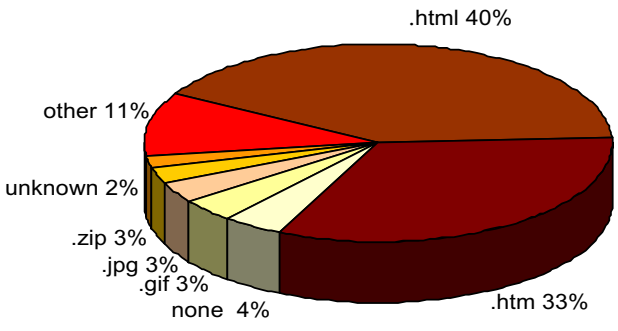

Figure 8: File type distribution

3.9 Languages Used

We built a language detector to identify language used in each page. From Figure 9, it is very surprised that English is major language used in Thai web. Number of pages that are written in English is around 66% of total number of pages. Almost all domains, Thai pages contribute less than 50%; except in `go.th` and `mi.th` domains.

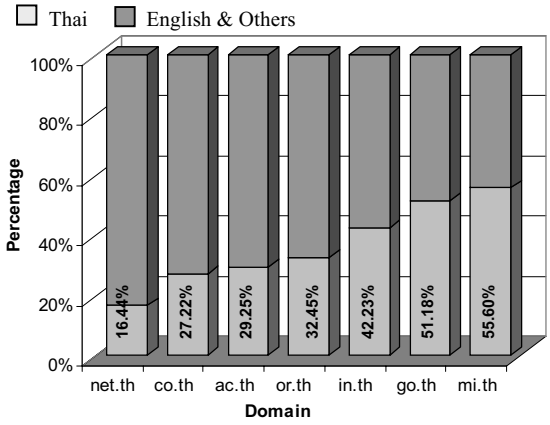

Figure 9: Languages used in each domain

4 Conclusions and Future Work

Quantitative measurement and statistics of Thai web are presented and analyzed. We plan to collect the Thai web pages every month to track the growth rate. Furthermore, there are many Thai servers that are registered under foreign domain. (e.g. `.com`, `.net`, `.org`, etc.) Recently investigation shows that there are 761 Thai servers that are registered under foreign domains [Upatisong00]. They would be included in the future measurement.

References

[daSilva98] A. da Silva, E. Veloso, P. Golghe, B. Ribeiro-Neto, A. Laender, and N. Ziviani. *CoBWeb — A Crawler for the Brazilian Web.* Proceedings of the String Processing and Information Retrieval Symposium & International Workshop on Groupware, 1998.

[Eichmann94] D. Eichmann. *The RBSE Spider- Balancing Effective Search Against Web Load*, Proceedings of the First International Conference on World Wide Web , pages 113-120, Geneva, Switzerland, May 1994.

[IIRC00] Internet Information Resource Center. *Domain counts and Host counts for .TH domain*. May 2000. <http://ntl.nectec.or.th/internet/domainname/WEB/>.

[Koanantakool94] T. Koanantakool, T. Tansethi, M. Kulatumyotin, *THAISARN: The Internet of Thailand*, July 1994. <http://ntl.nectec.or.th/thaisarn/thaisarn-body.html>.

[Koster93] M. Koster. *Database of Web Robots, Overview*. 1993. <http://info.webcrawler.com/mak/projects/robots/active/html/index.html>

[Pinkerton94] B. Pinkerton. *Finding What People Want: Experience with the WebCrawler*. Proceedings of the Second International Conference on World Wide Web, October 1994.

[Roehrl99] A. Roehrl., M.Frey, and R. Roehrl. *World Wide Web Robot for Extreme Datamining with Swiss-Tx Supercomputers*. Interim Report IIASA IR-99-20, 1999.

[Sanguanpong98]. S. Sanguanpong and S. Warangrit. *NontriSearch: Search Engine for Campus Network*. National Computer Science and Engineering Conference, Bangkok, Thailand, 1998.

[Upatisong00] V. Upatisong, *Electronics Personal Communications*, February 2000.