Characterization of the Thai Hostgraph

Kulwadee Somboonviwat
Dept. of Info. and Comm. Engineering
The University of Tokyo
kulwadee@tkl.iis.u-tokyo.ac.jp

Shinji Suzuki
Institute of Industrial Science
The University of Tokyo
suzuki@tkl.iis.u-tokyo.ac.jp

Masashi Toyoda
Institute of Industrial Science
The University of Tokyo
toyoda@tkl.iis.u-tokyo.ac.jp

Masaru Kitsuregawa
Institute of Industrial Science
The University of Tokyo
kitsure@tkl.iis.u-tokyo.ac.jp

ABSTRACT

The Web of a country or the national Web is a set of web pages related to a specific country. Understanding in the graph structure of the national Web provides invaluable insights for the development of algorithms and localized search services targeting for a specific country. Many empirical studies on the graph structure of the national Webs have been done at the level of individual web pages. However, in reality, the Web information is being organized into a hierarchically nested structure, called a domain name system. The domain name based hierarchical structure adds the intermediate levels of entities and administrative control to the Web. To better understand the characteristics and ecology of the national Web, it is necessary to also understand its graph structure at a more abstract level.

In this paper we put our attention to the graph structure of the Web at the level of interconnection between hosts in the Thai Web. The *hostgraph* is a directed graph with a node corresponding to a host and a directed weighted edge corresponding to the number of links between a pair of hosts. We report various graphical properties of the Thai hostgraph based on a snapshot of the Thai Web obtained in January 2007. For each empirical result, we carefully interpret its implications and discuss how to put it into practical use. We also give an example application of the hostgraph i.e. mining web community from the Thai hostgraph.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; H.3.5 [**Information Storage and Retrieval**]: Digital Libraries—*Collection*

General Terms

Experimentation, Measurement

Keywords

Web Characterization, Web Measurement

1. INTRODUCTION

In recent years, there have been several empirical studies on the characteristics of the Web of many different countries. The Web of a country or the national Web is a set of web pages related to a specific country. Understanding in the graph structure of the national Web provides valuable insights for the development of better search and navigation algorithms for localized search services targeting a specific country.

Many empirical studies on a graphical structure of the national Web (e.g. [5, 8, 9]) have been done at the level of individual web pages by using the *webgraph* model. However, in reality, the Web information is being organized into a hierarchically nested structure, called a domain name system. The domain name based hierarchical structure adds intermediate layers of entities and administrative control to the Web. To better understand the characteristics and ecology of the national Web, it is also necessary to understand its graph structure at a more abstract level i.e. hosts and domains levels. [4] have proposed a notion of the *hostgraph*. According to [4], the hostgraph is a directed graph where each node represents a web host and each directed edge represents the hyperlinks from web pages on the source host to web pages on the target host. The weight of the edge is equal to the number of such hyperlinks. [4] also raises many reasons for studying the hostgraph.

In this paper we study the graph structure of the Thai Web at the level of web hosts. The Thai hostgraph is generated from an archive of the Thai web crawled during January 2007. We have implemented a special-purpose web crawler for periodically creating snapshots of the Thai Web. Starting from a few popular Thai web portals, our crawler downloads Thai web pages and navigates the Web using a set of specialized crawling strategies. In the crawling process, the crawling strategies guide our web crawler to focus on the regions of the Web where there are many Thai web pages while at the same time keeping away from the regions of the Web where Thai web pages are scarce resources. We will give a formal definition of the Thai Web and discuss in more detail about Thai web crawling and the Thai Web dataset in Section 3.

Based on the Thai Web dataset, we have generated a Thai hostgraph consisting of 1.2 million nodes and 2.9 million weighted edges. We conduct statistical analysis on the characteristics of the Thai hosts and linkage structure of the Thai hostgraph. The main results include the power-law distribution of the number of pages per host, the power-law degree distribution, and the macroscopic linkage structure of the Thai hostgraph. Our analysis results uncover several unknown properties hidden in the link structure of the Thai Web. The implications of our results suggest potential opportunities for: (1) designing of a more robust web-spam detection algorithm, (2) designing of a more efficient web crawling algorithm, (3) improving website accessibility and navigation.

In the last part of the paper, we give an example of the utility of the Thai hostgraph i.e. finding a set of web hosts relating to a topic. The set of topically focused web hosts is extracted by applying a web community extraction algorithm [13] on the Thai hostgraph. We show a sample of the extracted communities of web hosts, and demonstrate a potential application of the hostgraph of a country.

The rest of the paper is organized as follows. Section 2 reviews the related literatures. Section 3 describes the Thai Web dataset. Section 4 presents characteristics of the Thai hosts. Section 5 reports various graphical properties of the Thai hostgraph. Section 6 gives an example application of the hostgraph: identification of web communities. Finally, Section 7 concludes the paper and discusses directions for further works.

2. RELATED WORK

Early works on the study of the Web as a graph include for example [1, 3, 6]. [1, 3] analyzes the Web graph of the University of NotreDame (300K nodes and 1.4M edges). [6] studies various graphical properties of the Web using two large datasets from AltaVista crawls (with more than 200 million nodes, and 1.5 billion links). These studies have consistently reported on emerging properties of the webgraph. One of the most remarkable result is the ubiquity of the power-law distributions such as power-law of host sizes, power-law of webgraph connectivity, power-law of connected components size, and power-law of PageRank score.

[6] analyzes the connected components and runs random breadth-first search traversal experiments on the webgraph induced from the AltaVista crawls. The interpretation of their results reveals a macroscopic linkage structure of the Web which can be depicted by a bow-tie like structure. The implication of the uncovered bow-tie structure of the Web provides a striking new insight into other aspects of the Web's graphical properties i.e. because there is a disconnected component in the bow-tie structure, it follows that the average and maximal diameter of the Web are infinite. Recent study on the properties of large webgraph has been conducted by [7]. They comprehensively analyzed the link structure of the webgraph, generated from a crawl of the Stanford WebBase Project[1]. They observed that the graphical properties of the WebBase sample (crawled in year 2001) are slightly different from the older sample studied in the prior works. Note that, the WebBase crawl dataset consists of approximately 200M pages and 1.4 billion edges.

The notion of the hostgraph have been firstly proposed by [4]. According to [4], the hostgraph is a directed graph where each node represents a web host and each directed edge represents the hyperlinks from web pages on the source host to web pages on the target host. The weight of the edge is equal to the number of such hyperlinks. [4] raises many convincing reasons for studying the hostgraph, and also demonstrates its practicality.

The studies of the webgraph of a country have been done by several countries such as [2, 5, 8, 9]. These studies reveal many interesting characteristics of the Web subgraphs pertaining to specific countries. Many statistical analyses have been done on the national webgraphs e.g. the African Web [5], the Web of Spain [2], the China Web [9], the Korean Web [8]. However, there is less work on the hostgraph of a country. To the best of our knowledge, only [2] and [9] provide the link structure analysis results of the national hostgraphs of Spain and China respectively.

[10] presents quantitative measurements and analyses on various properties of the Thai Web. Their dataset consists of 700K web pages downloaded from 8K web servers registered under '.th' domain on March 2000. The study in [10] only provides analysis results relating to the content and technological usage in the Thai Web. In this paper we empirically study the link structure of the Thai Web at the level of web hosts (i.e. the Thai hostgraph). The Thai hostgraph used in our study consists of 1.2 million nodes and 2.9 million weighted edges. In the following section, we will discuss about the definition of the Thai Web, the implementation of the crawler for Thai Web crawling, and the Thai Web dataset obtained by our crawler.

3. THAI WEB DATASET

In the World Wide Web, unlike in the real world, country border does not exist. As a result, the first challenge in comprehensively crawling the Web of a country lies in how to determine whether a web page belongs to the country or not. To overcome this challenge, most national web archiving projects use the country-code top-level domains (ccTLDs) and/or physical locations of the web servers assigned by IP addresses as the criteria of web page selection. According to the observations in [11, 12], many web pages written in the Thai language are residing outside the '.th' top-level domain of Thailand. Therefore, these two criteria is not appropriate for Thai web crawling because it would result in low coverage of the Thai Web. We propose the following set of criteria as a more appropriate strategy for crawling the Thai Web.

(1) Top-level domain of the web page is '.th'.

(2) IP address of its web server is physically assigned in 'Thailand'.

(3) Language of the web page is 'Thai'.

The first criterion can be easily implemented by adding a predicate function to check the value of the top-level domain of each URL before adding it into the URL queue of a crawler. For the second criterion, we need to check a geographical location of an IP address of each web server. We use an open source GeoIP API [2] to translate an IP address of a web server to its physical location (i.e. the country). The third criterion states that a web page should be included into the dataset if it is written in Thai regardless of its top-level domain. We implement the third criterion us-

[1] The Stanford WebBase Project:
http://dbpubs.stanford.edu:8091/∼testbed/doc2/WebBase/

[2] GeoIP API is available at
http://www.maxmind.com/app/geolitecountry.

ing a modified version of a language-specific web crawling method proposed in [11, 12]. In January 2007, we crawled the Thai Web using a few number of popular websites and web portals in Thailand as a start seed URLs. The resulting Thai Web dataset consists of 551,233 html pages, and about 51% of these crawled pages (280,429 html pages) are written in Thai language.

4. CHARACTERISTICS OF THAI HOSTS

We define a host as a set of pages sharing the server part of a canonical form of the URL For example, a server part of a URL address http://www.asite.co.th:80/index.html is www.asite.co.th:80. Based on the above definition, there are 1,214,457 hosts in the hostgraph generated from our Thai Web dataset (including both crawled and uncrawled nodes). The number of hosts under '.th' domain is 27,668. The average numbers of pages per host are 5 pages per host (all hosts) and 9 pages per host (only host under '.th' domain) respectively.

Fig. 1 shows two plots of the distribution of the number of pages per host. The graph in Fig. 1(a) is the distribution of all hosts in the dataset. In Fig. 1(b), we plot the distribution only for the hosts under the Thailand national domain name. From the figure, it can be seen that both distributions are very skewed with the middle parts of the graphs fitted the power-law distribution.

In the case of all hosts (Fig. 1(a)), the distribution of the number of pages per hosts follows the power-law when the number of pages per host is in the range of [20,1000]. The approximated power-law exponent is 2.00. When the number of pages per host is between 199 and 204, we can see some outliers in the power-law plot of Fig. 1(a). After examining the URLs and the content of web pages corresponding to these outliers, we found that most web pages are spams generated automatically by machine. They have very similar patterns of URL addresses and also very similar content pattern.

In the case of the hosts under '.th' domain (Fig. 1(b)), the host sizes distribution follows the power-law when the number of pages per host is in the range of [20,1000] and the best-fit power-law exponent is approximately 1.68. Considering the power-law plots in (a) and (b), we find that while (a) is better fit to the power-law no anomalous outliers is observed in (b). This suggests that the hosts under the '.th' national domain are more spam-free than the hosts outside the '.th' domain.

By comparing the power-law exponent of the host sizes distribution of the Thai Web with other countries, we can estimate the relative number of large hosts in each national Web. Specifically, the larger the power-law exponent, the larger there is the relative number of large hosts. The power-law exponents of host sizes distributions are equal to 1.14 for Spain [2], 1.74 for China [9], 2.00 for Thailand, and 2.3 for South Korea [8].

5. THE THAI HOSTGRAPH

A *webgraph* [6] is a directed graph where each node represents a web page and each directed edge represents the hyperlink from a source web page to a destination web page. A *hostgraph* [4] is a directed graph where each node represents a web host and each directed edge represents the hyperlinks from web pages on the source host to web pages on the destination host. The weight of the edge is equal to the number of such hyperlinks.

Table 1: Properties of underlying Thai webgraph

Vertices (web pages) in the webgraph	5,785,349
Directed edges in the webgraph (millions)	12

Table 2: Properties of the Thai hostgraph

Vertices (hosts) in the hostgraph	1,214,457
Hosts in '.th' domain	27,668
Directed edges in the hostgraph	2,904,632
Inter-host hyperlinks (sum of edge weights)	8,878,268
Percentage of intra-host hyperlinks	60.4%

Based on the aboved definition of a hostgraph, we have generated the hostgraph using the Thai Web dataset (described in Section 3). Table 2 and Table 1 show properties of the Thai hostgraph and the underlying webgraph respectively. Of all 5.7 million hyperlinks in the Thai webgraph, about 60% of those hyperlinks are intra-host hyperlinks (i.e. hyperlinks between web pages on the same host). According to Table 2, our Thai hostgraph is sparse, it consists of 1.2 million nodes and 2.9 million directed edges. There are 27K nodes in the hostgraph corresponding to hosts under '.th' domain.

In the following subsections, we will present some characteristics of the Thai hostgraph extracted from our dataset.

5.1 In-degree and Out-degree Distributions

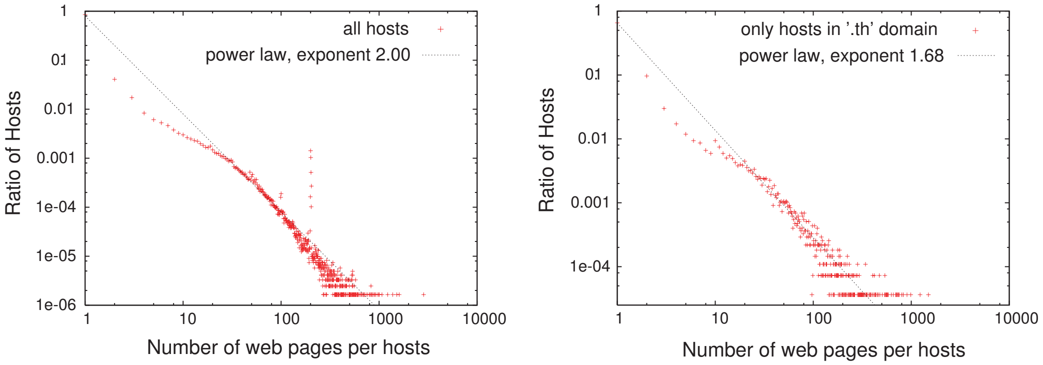
The distributions of the weighted in-degree and out-degree of hosts in the Thai hostgraph exhibit the power-law distribution with the exponent of 1.85 and 1.44 respectively, as shown in Fig. 2. In Fig. 3, we plot the weighted in-degree and out-degree distributions of the hosts under '.th' domain. While the weighted in-degree distribution of '.th' hosts indicates the power-law distribution with exponent 1.45, the weighted out-degree distribution of '.th' hosts does not fit with the power-law distribution (we obtained the power-law exponent whose value is less than 1 when trying to fit the log-log plot with the power-law). Among the hosts with high weighted in-degree, we observe mostly Thai portals and news websites. The hosts with high weighted out-degree are mostly Web directories and blog service websites.

For comparison, we show power-law exponents values of the hostgraphs of different Web regions as reported in previous published works in Table 3.

5.2 Strongly Connected Components

In graph theory, a *connected component* of an undirected graph is a set of nodes such that for any pair of nodes u and v, there exists a path from u to v. A *strongly connected component (SCC)* of a directed graph G is a set of nodes S such that for every pair of nodes $u, v \in S$, there exists a *path* from u to v and from v to u. A *weakly connected component (WCC)* of a directed graph G is a set of nodes W where W is a connected component of the undirected graph obtained by ignoring the directions of all edges in G.

We can extract 9,457 WCC components and 1,138,627 SCC components from the Thai hostgraph (total number of nodes = 1,214,457). The largest WCC component contains 98.7% of nodes in the Thai hostgraph. The largest SCC



(a) All hosts.

(b) '.th' hosts.

Figure 1: Distribution of the number of pages per host

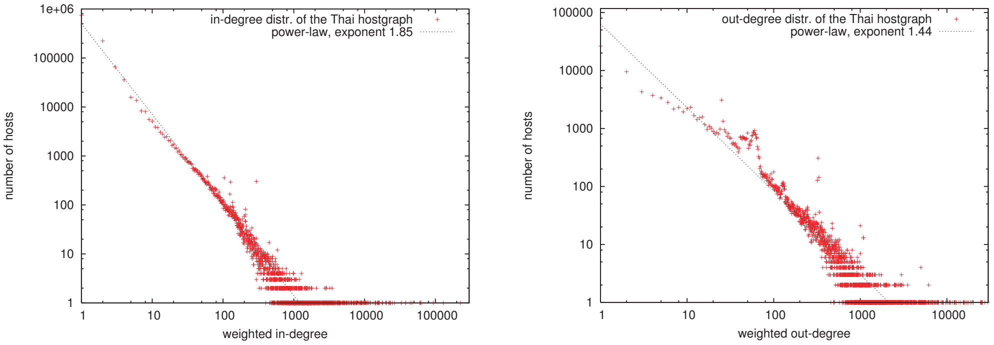


Figure 2: Degree Distribution of the Thai hostgraph

Table 3: Weighted in-degree and out-degree of the hostgraphs from different regions of the Web

Web region	weighted in-degree	weighted out-degree
Global Web [4]	1.62	1.67
China [9]	1.40	1.50
Spain [2]	1.82	1.34
Thailand	1.85	1.44

Table 4: Size of each component in the bow-tie structure of the Thai hostgraph

Component	Size(number of hosts)	Percentage[%]
MAIN	68891	5.7
IN	2415	0.2
OUT	892816	73.5
TENDRILS	234751	19.3
ISLANDS	15584	1.3

component consists of 68,891 nodes in the Thai hostgraph. The number of SCC component with single host is 1,137,100

The distribution of the sizes of SCC in the Thai hostgraph also indicate the power-law distribution with exponent 2.75. The log-log plot of the SCC sizes distribution is as shown in Fig. 4.

5.3 Large-scale Link Structure

According to [6], the large-scale link structure of the Web can be depicted as a bow-tie. The bow-tie structure consists of five components as follows

- *MAIN*: consists of web pages in the largest strongly connected component (SCC) in the graph.

- *IN*: consists of web pages that can reach the MAIN but cannot be reached from the MAIN.

- *OUT*: consists of web pages that can be reached from the MAIN but cannot reach any pages in MAIN.

- *TENDRILS*: consists of web pages that can be reached from IN and those that can only reach to OUT.

- *ISLANDS*: consists of web pages outside the largest weakly connected component in the graph.

The distribution of the hosts in each component is given in Table 4. And, we also show the distribution of the domains in which the hosts on each component are in Table 5. From Table 4, we found that most of the Thai hosts are located in the OUT component and there is a small number of hosts in the IN component. The reason for very small IN component might be caused by the limitation of the crawling. (i.e. because the hosts in the IN component can only be reached if the URL addresses of these hosts are known beforehand.)

6. MINING WEB COMMUNITY FROM THE THAI HOSTGRAPH

In this section, we give an example of the utility of the Thai hostgraph: finding a set of web hosts relating to a

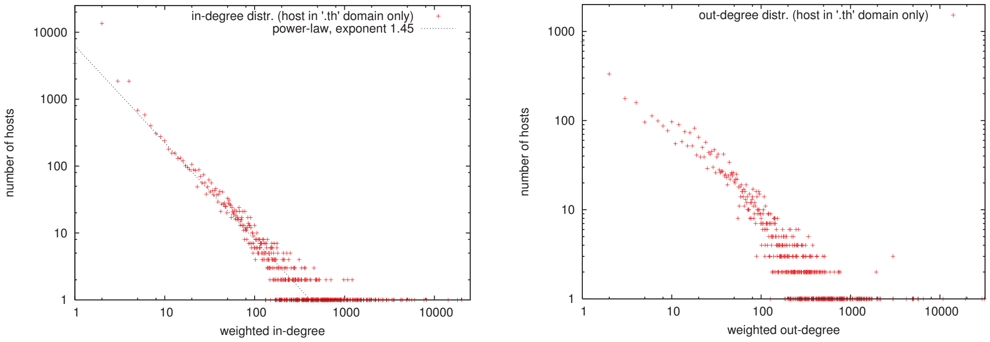


Figure 3: Degree Distribution of the Thai hostgraph (only hosts in the '.th' domain)

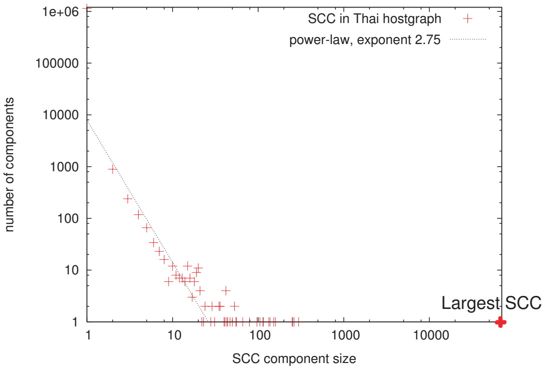


Figure 4: Distribution of SCC component size in the Thai hostgraph

Table 5: Distribution of the domains in each components

Comp.	'.th'	'.com'	'.net'	other	total
MAIN	4097	49678	3100	12016	68891
	5.95%	72.11%	4.50%	17.44%	100.0%
IN	57	1299	65	994	2415
	2.36%	53.79%	2.69%	41.16%	100.0%
OUT	23349	281571	35137	552759	892816
	2.61%	31.54%	3.94%	61.91%	100.0%
TENDR.	68	5776	452	228455	234751
	0.02%	2.46%	0.20%	97.32%	100.00%
ISLAND	97	5589	336	9562	15584
	0.62%	35.86%	2.16%	61.36%	100.0%

topic. The set of topically focused web hosts is extracted by applying a web community extraction algorithm proposed in [13] on the Thai hostgraph. We show an example of the extracted communities of web hosts in Table 6.

All URLs listed in Table 6 are homepages of news and televisions in Thailand. This example demonstrate one of a potential applications of the national hostgraph.

7. CONCLUSION

In this paper, we have comprehensively studied the hostgraph of Thai Web. We reported various graphical properties of the Thai hostgraph based on a snapshot of the Thai Web obtained in January 2007. For each empirical result, we carefully interpret its implications and discuss how to put it into practical use. We also give an example application of the hostgraph: mining web community from the Thai hostgraph. The statistics of link structure of the Thai hostgraph presented in this paper can be used in the design of more efficient web crawling and searching algorithms. For the future work, we plan to study the evolution of the link structure of the Thai Web using our Thai Web archives. We also plan to do content analysis on our Thai web snapshots.

8. REFERENCES

[1] R. Albert, H. Jeong, and A. Barabasi. The diameter of the world wide web. *Nature*, 401:130, 1999.

[2] R. Baeza-Yates, C. Castillo, and V. Lopez. Characteristics of the web of spain. *International Journal of Scientometrics, Informetrics and Bibliometrics*, 9(1), 2005.

[3] A. Barabsi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[4] K. Bharat, B.-W. Chang, M. R. Henzinger, and M. Ruhl. Who links to whom: Mining linkage between web sites. In *Proc. of the 2001 IEEE Int'l Conf. on Data Mining (ICDM'01)*, pages 51–58, 2001.

[5] P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Structural properties of the african web. In *Poster Proc. of the 11th Int'l Conf. on World Wide Web (WWW'02)*, 2002.

[6] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. L. Wiener. Graph structure in the web. *Computer Networks*, 33(1–6):309–320, 2000.

[7] D. Donato, L. Laura, S. Leonardi, and S. Millozzi.

Table 6: Community of News&Media extracted from the Thai hostgraph

http://www.bangkokpost.com/
http://www.banmuang.co.th/
http://www.manager.co.th/
http://www.matichon.co.th/
http://www.naewna.com/
http://www.nationmultimedia.com/
http://www.siamrath.co.th/
http://www.thannews.th.com/
http://www.tv5.co.th/
http://www.ubctv.com/
http://www.bangkokbiznews.com/
http://www.mcot.or.th/
http://www.posttoday.com/
http://www.siamturakij.com/
http://www.ch7.com/
http://www.dailynews.co.th/
http://www.itv.co.th/
http://www.thaipost.net/
http://www.thairath.co.th/
http://www.thaitv3.com/

The web as a graph: How far we are. *ACM Trans. Inter. Tech.*, 7(1):4, 2007.

[8] I. K. Han, S. H. Lee, and S. Lee. Graph structure of the korea web. In *Proc. of the 12th Int'l Conf. on Database Systems for Advanced Applications (DASFAA'07)*, pages 930–935, 2007.

[9] G. Liu, Y. Yu, J. Han, and G.-R. Xue. China web graph measurements and evolution. In *Proc. of the 7th Asia Pacific Web Conference (APWeb'05)*, pages 668–679, 2005.

[10] S. Sanguanpong, P. Piamsa-nga, Y. Poovarawan, and S. Warangrit. Measuring and analysis of the thai world wide web. In *Proc. of the Asia Pacific Advance Network conference*, pages 225–330, 2000.

[11] K. Somboonviwat, T. Tamura, and M. Kitsuregawa. Finding thai web pages in foreign web spaces. In *ICDE Workshops*, page 135, 2006.

[12] T. Tamura, K. Somboonviwat, and M. Kitsuregawa. A method for language-specific web crawling and its evaluation. *Systems and Computers in Japan*, 38(2):10–20, 2007.

[13] M. Toyoda and M. Kitsuregawa. Creating a web community chart for navigating related communities. In *Proc. of the 12th ACM conference on Hypertext and Hypermedia (HYPERTEXT '01)*, pages 103–112, 2001.