Web Structure in 2005

Yu Hirate, Shin Kato*, and Hayato Yamana

Dept. of Computer Science, Waseda University, 3-4-1 Okubo Shinjuku-ku Tokyo, Japan
{hirate,kato,yamana}@yama.info.waseda.ac.jp

Abstract. The estimated number of static web pages in Oct 2005 was over 20.3 billion, which was determined by multiplying the average number of pages per web server based on the results of three previous studies, 200 pages, by the estimated number of web servers on the Internet, 101.4 million. However, based on the analysis of 8.5 billion web pages that we crawled by Oct. 2005, we estimate the total number of web pages to be 53.7 billion. This is because the number of dynamic web pages has increased rapidly in recent years. We also analyzed the web structure using 3 billion of the 8.5 billion web pages that we have crawled. Our results indicate that the size of the "CORE," the central component of the bow tie structure, has increased in recent years, especially in the Chinese and Japanese web.

1 Introduction

As of Oct. 2005, the number of static web pages was estimated at over 20.3 billion. This number was calculated by multiplying the estimated average number of web pages per web server, 200[1][2][3], by the number of web servers in Oct. 2005, 101,435,253[4]. However, based on our crawling status by Oct. 2005[5], we estimate the number of web pages, including both static and dynamic web pages, in Oct. 2005 to be about 53.7 billion. This discrepancy may be due to the increase in number of dynamic web pages generated by CGI, *etc.*

In 1999, Broder *et al.* analyzed the graph structure of the web, called the web structure, from the set of web pages crawled in that year[6]. In this previous report, about 90% of web pages belonged to connected components, and about 28% of web pages belonged to SCC (=strongly connected components)[6]. Inspired by Broder's web structure, several researchers investigated the web structure based on their crawling web data. For example, in 2003, Lie *et al.* analyzed the structure based on the set of web pages crawled from China. They reported that the percentage of SCC was much larger than that in Broder's web structure[8]. However, there have been no analyses of the web structure based on recent web pages from all over the world. Here, we report the web structure computed from 3 billion web pages crawled between Jan. 2004 and Oct. 2005.

The remainder of this paper is organized as follows. In section 2, we describe the e-Society Project[5] funded by the Japanese government. We analyzed the

[*] Currently working at Mitsubishi Electric Corporation.

W. Aiello et al. (Eds.): WAW 2006, LNCS 4936, pp. 36–46, 2008.

web structure based on the data of web pages crawled by the e-Society Project. In section 3, we report the estimated number of web pages. In section 4, we show examples of related work with respect to analysis of the web structure. In section 5, we report the results of analysis of the web structure based on our crawled web pages. Section 6 presents a summary of our work.

2 The e-Society Project[5]

The e-Society project "Technologies for the Knowledge Discovery from the Internet" is one of the projects of the Ministry of Education, Culture, Sports, Science, and Technology, Japan. The project contractor is Waseda University. The project aims (1) to gather all web pages in the world efficiently and (2) to discover some type of knowledge by applying data mining techniques. To achieve these goals, we are now gathering web pages from all over the world. As described in section 1, we used the data of pages crawled as part of this project to analyze the web structure in 2005.

2.1 Crawling Status

We began gathering web pages in Jan. 2004 with 30 CPUs in 3 different locations in Japan. We added 20 CPUs in Jan. 2005 and another 30 CPUs in Oct. 2005. Currently, our crawling system has the capability to gather up to 35 million web pages per day. By Oct. 2006, we had gathered over 14 billion web pages from all over the world.

3 How Many Web Pages Are There?

As the most basic analysis, we estimated the number of web pages on the web. Conventional research indicates that each web host has an average of about 200 web pages[1][2][3]. Netcraft, a company from the UK, investigates the number of web servers on the whole web and publishes their results every month on their home page. According to the Netcraft report of Nov. 2005[4], the number of web servers was estimated as 101,435,253. By multiplying these two numbers -*i.e.*, $200 \times 101,435,253$- we can estimate the total number of web pages as about 20.3 billion.

However, by Oct. 2005, we had gathered 8,507,237,370 web pages from 16,035, 801 web servers, indicating that the average number of web pages per host is about $8,507,237,370/16,035,801 \simeq 530$. Multiplying this figure by the number of web servers reported by Netcraft yields an estimate of the total number of web pages all over the world of about 53.7 billion. This discrepancy may be due to the recent rapid increase in number of dynamic web pages, such as CGI pages based on databases, Blogs, Portal Sites, and EC sites.

Note that this analysis differs from estimating the number of web pages indexed by search engines. Bharat *et al.*[9], Henzinger *et al.*[10] Vaughan *et al.*[11] and Bar-Yossef *et al.*[12] investigated the relative size of several search engines' indexes in 1997, 1998, 2004 and 2006, respectively. These works focus on the relative size of indexed web pages by different search engines. On the other hand, our analysis estimates the size of actual web, regardless of whether each web page is indexed or not.

4 Related Works

In this section, we present related work with respect to web structure[6].

4.1 Applying Graph Theory to Web Link Structure

The web has a hyperlinked structure, which was first introduced by Broder *et al.* [6]. When we consider pages as vertexes and hyperlinks as edges, then it is possible to regard the web link structure as a directed graph. Broder *et al.* focused on this property, and analyzed the web link structure from the viewpoint of graph theory[6]. Inspired by their approach, several researchers analyzed the web structure based on their own web data. The remainder of this section describes these conventional studies.

4.2 Graph Structure in the Web[6]

Broder *et al.* analyzed the whole web structure in 1999 based on 200 million web pages with 1,500 million hyperlinks[6]. They reported that about 90% of web pages belong to connected components and the structure is similar in shape to a bow tie, as shown in Fig. 1.

Broder *et al.* defined four types of component in the web structure, IN, CORE, OUT, and TENDRILS[6]:

- CORE is defined as SCC(=Strongly Connected Component).
- IN is defined as the set of web pages that have paths to SCC but do not have paths from SCC.
- OUT is defined as the set of web pages that have paths from SCC but do not have paths to SCC.
- TENDRIL is defined as the set of web pages that do not have paths either to or from SCC.

As shown in Table 1, the web structure in 1999 consisted of 28% SCC and 21% IN, OUT, and TENDRILS.

4.3 Structural Properties of the African Web[7]

Boldi *et al.* analyzed the African web structure in 2002. The dataset consisted of 2 million web pages gathered from 2,500 African hosts. They reported that the African web differed in shape from the bow tie structure of the web as a whole. As shown in Fig. 2, the African web structure contained multiple OUT components, but had no IN components. However, the dataset of the African web structure was small compared to those of other web structures, and this may have been responsible for the difference in shape.

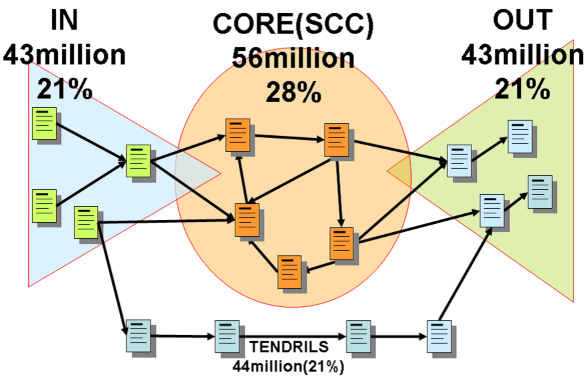

Fig. 1. Bow-tie Structure of the Web in 1999[6]

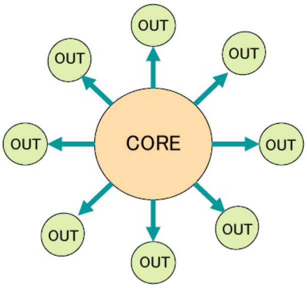

Fig. 2. African Web Structure in 2002[7]

4.4 China Web Graph Measurements and Evolution[8]

Lie *et al.* analyzed the Chinese web structure in 2003 based on 140 million web pages with 4,300 million hyperlinks[8]. They reported that the Chinese web structure was bow tie-shaped, as shown in Fig. 3. However, as shown in Table 1, the percentage of CORE in the Chinese web structure in 2003 was much larger than that of in Broder's whole web structure in 1999. The authors concluded that the large percentage of CORE in the Chinese web structure was a phenomenon specific to the Chinese web.

5 Web Structure in 2005

As we described in section 4, conventional studies have revealed different properties of the web structure. Broder's web structure was based on the web data in 1999. Boldi's African web structure and Lie's Chinese web structure were based on parts of the whole web. To determine the current shape of the web structure, an updated web structure is needed. We analyzed the structure based on data

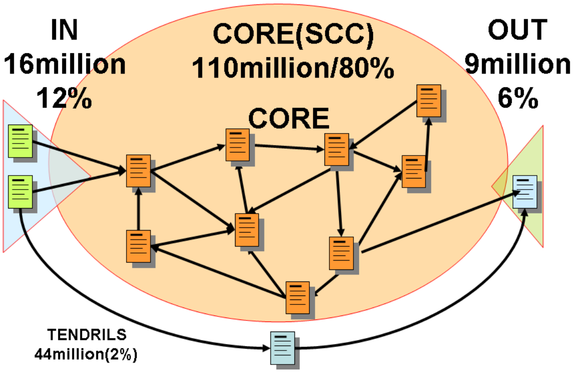

Fig. 3. Chinese Web Structure in 2003[8]

Table 1. The percentages of components of bow-tie structures[6][8]

WebGraph	CORE(SCC)	IN	OUT	TENDRILS	DISCONNECTED
Web in 1999[6]	56 million (28%)	43 million (21%)	43 million (21%)	44 million (21%)	17 million (8%)
Chinese Web in 2003[8]	112 million (80%)	16.5 million (12%)	9 million (6%)	1 million (0.7%)	1 million (0.7%)

consisting of web pages gathered as part of the e-Society project. In this section, we report the results for the current web structure.

We analyzed the whole web structure, web structures by TLD (=Top Level Domain), and web structures by language based on 3,207,736,427 web pages[1] gathered between Jan. 2004 and Jul. 2005. These web pages were gathered from all over the world and their languages were detected automatically using the Basis Technology Rosette Language Identifier[13].

The reminder of this section is as follows. In section 5.1, we introduce our analytical strategy for computation, *i.e.*, host level reduction. In section 5.2, we describe the properties of our dataset. Then, we report the results of the whole web structure, web structures by TLD, and web structures by language in sections 5.3, 5.4, and 5.5, respectively.

5.1 Host Level Reduction

Our analysis was based on host level analysis as described below:

- Pages in the same host are considered as one vertex.
- Hyperlinks to other hosts are considered as edges.

[1] Our crawler gathers web pages from the top page of each host, and follows links up to 15 stratums.

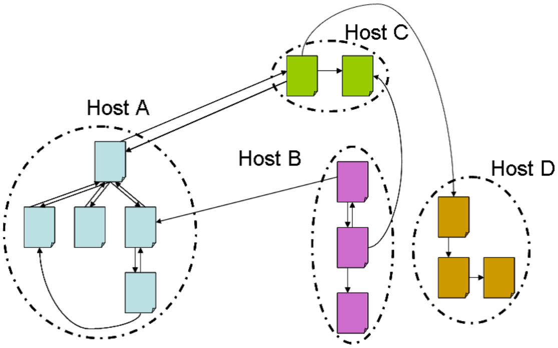

Fig. 4. Raw Link Data

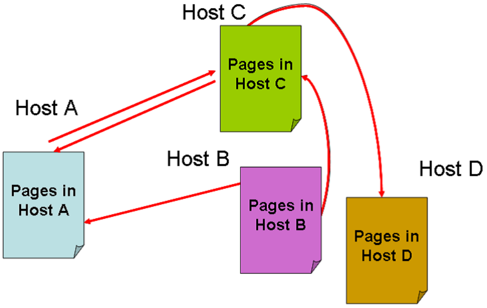

Fig. 5. Preprocessed Link Data

To analyze the host level web structure, we applied the following preprocessing steps to a raw link dataset before computing the web structure.

1. Extract hosts from a raw link dataset to generate a host list.
2. Group web pages in the same host.
3. Extract links that connect two different hosts, called inter-host links, from the raw link dataset.

We call these preprocessing steps "host level reduction." For example, when we have a raw link dataset as shown in Fig. 4, preprocessed data will be as shown in Fig. 5.

5.2 Dataset Properties

We used the data of web pages crawled as part of the e-Society project. The TLD distribution and language distribution of the dataset are shown in Fig. 6 and Fig. 7, respectively. Similar to the actual web, our data collection consisted of a large proportion of ".com" domains (Fig. 6) and web pages in English (Fig. 7).

Table 2. Components of Whole Web Structure in 2005

	CORE	IN	OUT	Other
Number of hosts	624,173	147,794	621,788	119,770
Percentage of hosts	41%	10%	S 41%	8%
Number of pages	2,102,971,321	633,530,035	346,251,616	124,983,455
Percentage of pages	65%	20%	11%	4%

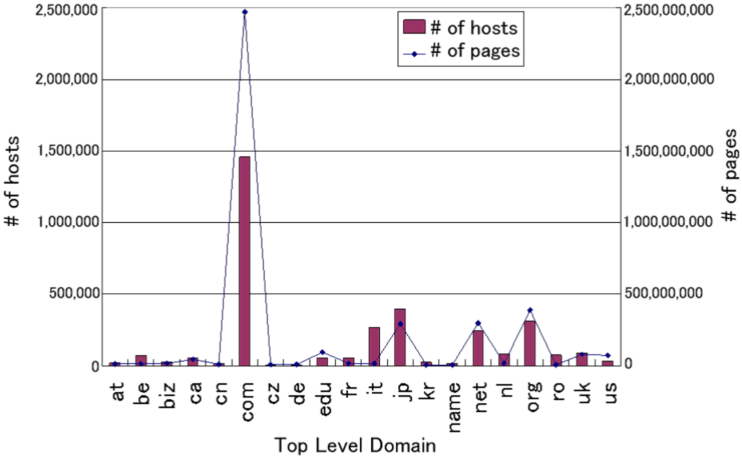

Fig. 6. TLD Distribution of the Dataset

However, as we gathered web pages from ".jp" and ".com" domain lists, our dataset was biased toward ".jp" and ".com" domains.

Our raw link dataset consisted of 3,208,139,905 web pages (vertexes) and 93,397,065,743 links (edges). After applying host level reduction preprocessing, the preprocessed link data consisted of 1,719,134 hosts (vertexes) and 91,084,879 inter-host links (edges).

In this analysis, we have not discarded duplicated web pages in a host nor among hosts. However, because of the host level reduction, which is described in section 5.1, web pages in the same host are considered as a vertex. Due to this, our result turns to be the same even if duplicated web pages in the same host are aggregated. On the other hand, when duplicated web pages exist among hosts, they are considered as multiple vertexes.

5.3 The Whole Web Structure in 2005

Fig. 8 and Table 2 show the results of the whole web structure in 2005. As shown in Fig. 8 and Table 2, the percentage of the CORE component had become larger than that in Broder's whole web structure in 1999. Although Lie *et al.* concluded

Table 3. Components of web structures by TLD in 2005

TLD	CORE	IN	OUT	Other
.com	53.65%	19.73%	22.25%	4.37%
.jp	26.46%	1.77%	71.32%	0.46%
.de	0.25%	0.05%	78.36%	21.34%
.edu	0.05%	0.00%	14.44%	85.51%
.fr	0.01%	0.02%	25.33%	74.63%
.it	0.11%	0.04%	0.04%	99.81%
.kr	0.00%	0.00%	1.09%	98.91%
.net	0.52%	0.17%	35.42%	63.89%
.org	0.61%	0.38%	64.25%	34.76%
.ru	0.77%	0.05%	0.49%	98.70%

that the increase in the CORE component percentage was a pattern specific to the Chinese web, this phenomenon was found not only in the Chinese web, but also in the whole web in the present study.

5.4 Web Structures by TLD

Table 3 shows the results of web structures by TLD. As shown in Table 3, there were no large CORE components in web structures of all TLD, excluding the .com and .jp domains. Even in the .jp domain web structure, the percentage of CORE component was smaller than that in Broder's whole web structure in 1999. These results indicate that the web cannot be divided with regard to TLD.

5.5 Web Structures by Language

Table 4 shows the results of web structure by language. As shown in Table 4, Chinese and Japanese language web structures have large percentages of CORE

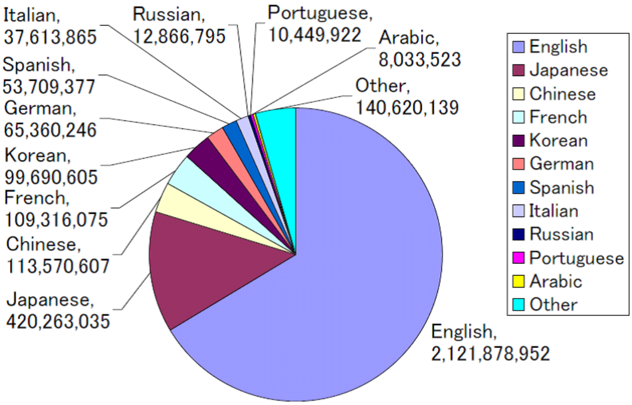

Fig. 7. Language Distribution of the Dataset

Table 4. Components of web structures by Language in 2005

TLD	CORE	IN	OUT	Other
Chinese	76.88%	9.98%	10.57%	2.57%
Japanese	71.05%	25.85%	2.54%	0.56%
English	66.90%	9.04%	16.44%	7.62%
Spanish	64.93%	5.30%	23.60%	6.16%
French	61.85%	9.23%	20.65%	8.27%
Arabic	61.43%	10.20%	18.59%	9.78%
Korea	54.32%	17.07%	19.36%	9.25%
Russian	35.76%	18.20%	18.35%	27.69%
Portuguese	26.60%	4.94%	42.18%	26.28%
German	26.61%	8.16%	42.18%	23.05%
Italian	23.67%	17.10%	29.54%	29.69%
Other	7.24%	1.98%	9.32%	81.47%

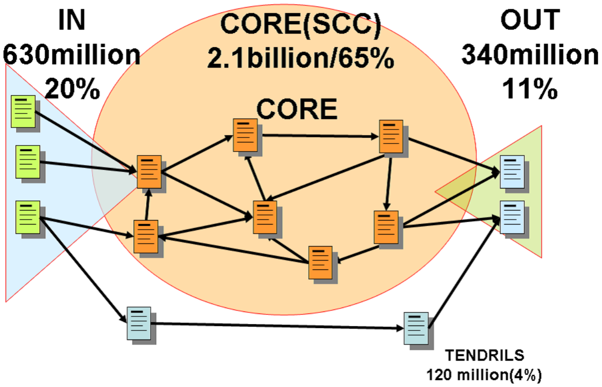

Fig. 8. Web structure in 2005

components. On the other hand, German and Italian language web structures have small percentages of CORE components. This may be because our web data collection had low coverage in the case of German and Italian language web pages. Note that in the case of the Chinese language web structure, the percentages of components were similar to those in Lie's Chinese web graph in 2003.

6 Conclusion

In this paper, we reported the properties of the web structure in 2005 based on 3.2 billion web pages crawled as part of the e-Society project. Compared to Broder's web structure in 1999, the percentage of the CORE component increased from 28% to 65%. Lie *et al.* concluded that the large percentage of CORE component is a phenomenon specific to the Chinese web structure. However, our analysis showed that the increase in CORE component has occurred in the whole web.

We also analyzed web structures by TLD and by language. By comparing the two types of web structure, we confirmed that the web cannot be divided by TLD, but can be divided by written language.

6.1 Future Work

Since our analysis is based on the crawled web pages which have been crawled by our original crawler following hyper-links from seed URLs, our dataset might loss some web pages which should be categorized as IN component. Then this might result in the decreased percentage of IN component to some extent. To solve this, one can use indexes of search engines, and internet archive, such as WebBase[14][15] to add the fraction of IN component to the dataset. As a future work, we update the web structure with solving such a IN component problem.

Acknowledgements

This research was funded in part by both "e-Society: the Comprehensive Development Foundation Software" of MEXT (Ministry of Education, Culture, Sports, Science, and Technology) and "21st Century COE Programs: ICT Productive Academia" of MEXT.

References

1. Lawrence, S., Giles, C.L.: Searching the World Wide Web. Science 280(5360), 98–100 (1998)
2. Lawrence, S., Giles, C.L.: Accessibility of Information on the Web. Nature, 400, 107–109 (1999)
3. Institute for Information and Communications Policy, Statistics Investigation Report for contents on the World-Wide Web (2004), `http://www.soumu.go.jp/iicp/chousakenkyu/seika/houkoku.html`
4. Netcraft: Web Server Survey (November 2006), `http://news.netcraft.com/archives/2006/11/01/november_2006_web_server_survey.html`
5. e-Society Project, `http://www.yama.info.waseda.ac.jp/~yamana/e-society/index_eng.htm`
6. Broder, A., Kumar, R., Maghoul, F., Raghavan, P., State, R., Tomkins, A., Wiener, J.: Graph structure in the web. In: Proc. of 9th World Wide Web Conf., pp. 309–320 (2000)
7. Boldi, P., Codenotti, B., Santini, M., Vigna, S.: Structural Properties of the African Web. In: Poster Proc. of 11th World Wide Web Conf., (2002)
8. Lie, G., Yu, Y., Han, J., Xue, G.: China web graph measurements and evolution. In: Zhang, Y., Tanaka, K., Yu, J.X., Wang, S., Li, M. (eds.) APWeb 2005. LNCS, vol. 3399, Springer, Heidelberg (2005)

9. Bharat, K., Broder, A.: A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines. Journal of Computer Networks and ISDN Systems 30(1-7), 379–388 (1998)
10. Henzinger, M., Heydon, A., Mitzenmacher, M., Najork, M.: Measuring Index Quality using Random Walks on the Web. In: Proc. of 8th World Wide Web Conf., pp. 213–225 (1999)
11. Vaughan, L., Thelwall, M.: Search Engine Coverage Bias: Evidence and Possible Causes. Journal of Information Processing and Management 40(4), 693–707 (2004)
12. Bar-Yossef, Z., Gurevich, M.: "Random Sampling from a Search Engine's Index. In: Proc. of 15th World Wide Web Conf., pp. 367–376 (2006)
13. Basis Technology Rosette Language Identifier, http://www.basistech.com/language-identification/
14. Hirai, H., Raghavan, S., G-Molina, H., Paepcke, A.: Webbase: A repository of the Web. In: Proc. of 9th World Wide Conf., pp. 277–293 (2000)
15. The Stanford WebBase Project, http://dbpubs.stanford.edu:8091/~testbed/doc2/WebBase/