한국 웹 그래프와 진화에 대한 연구

한 인 규[†] · 이 상 호[††]

요 약

웹 그래프에 대한 연구는 웹 문서의 효율적인 수집을 위하여 적용되는 알고리즘과, 커뮤니티의 검색 및 발견의 분야에 있어 매우 중요한 위치를 차지한다. 또한 웹 그래프의 연구에 있어 발견되는 웹의 현상들은 웹이 가지고 있는 특징들을 나타내며 웹 그래프의 진화를 연구함으로써 웹의 크기와 진화 프로세스를 예측할 수 있다. 본 논문에서는 약 1억 1천만 개의 노드와 약 27억 개의 노드를 가지는 한국 웹 그래프에 대한 연구를 수행한다. 먼저 한국 웹의 페이지들이 서로 얼마나 연결되어 있는가에 대한 접속도 연구를 수행한다. 한국 웹의 접속도는 bow-tie 모형으로 표현할 수 있다. 또한 Power Law 현상과 같은 한국 웹의 특징이 글로벌 웹과 어떤 차이가 있는지 분석한다. 한국 웹 그래프의 속성은 글로벌 웹과는 많은 차이를 보여주었다. 마지막으로 한국 웹 그래프의 진화에 대한 연구를 여러 가지 관점으로 수행한다.

키워드 : 웹 그래프, 멱함수, 웹진화

Graph Structure and Evolution of the Korea web

In Kyu Han[†] · Sang Ho Lee[††]

ABSTRACT

The study of the web graph yields valuable insight into web algorithms for crawling, searching and community discovery, and the sociological phenomena which characterize its evolution, also it is useful for understanding the evolution process of web graph and predicting the scale of the Web. In this paper, we report experimental results on properties of the Korea web graph with over 116 million pages and 2.7 billion links. We indicate to study the Korea web properties such as the power law phenomenon and then to analyze the similarity and difference between the global and Korea web graph. Our analysis reveals the Korea web graph have different properties compared with the global web graph from the structure to the evolution of the Web. Finally, a number of measurements of the evolution of the Korea web graph will be represented.

Key Words : web Graph, Power Law, web Evolution

1. Introduction

The World Wide Web has revolutionized the way we access information. The web has about a billion pages today, several billion links, and it is growing rapidly at the rate of 7.3 million pages a day [15]. This gigantic structure often makes it difficult for even the most technical users to find the best information available on a given topic. Recent studies suggest that despite its chaotic appearance, the web is a highly structured digraph, in a statistical sense. The link structure has useful information for measuring the importance of web pages. The link analysis algorithms such as PageRank play an important role in the performance of search engines. In order to keep up with the importance of web pages that are constantly changing at all times, it is important for search engines to accurately capture the web link structures. These features could be exploited to attain efficiency and comprehensiveness in web navigation.

The web can be modeled as a directed graph where each node represents a page, an edge, and a hyperlink. This directed graph is called a web graph. This graph approach aids in understanding the structure of the web at macroscopic as well as microscopic levels. An overview of applications of graph theory to the WWW appears in Hayes [16].

There are several reasons for developing an understanding of web graph. First, it is important to design crawl strategies on the Web. Analyzing the behavior of web algorithms that make use of link information, such as PageRank, is important, too. It can also be used to predict the evolution of web structures and develop better algorithms for discovering and organizing them.

※ 본 연구는 숭실대학교 교내연구비 지원으로 이루어졌음.
† 정 회 원 : 숭실대학교 대학원 컴퓨터학과 석사
†† 종신회원 : 숭실대학교 컴퓨터학부 교수
논문접수 : 2007년 1월 14일, 심사완료 : 2007년 3월 26일

We study various properties of the Korea web graph, including its degree distributions, connected components, and macroscopic structure. In addition, we studied the evolution of the Korea Web. We performed three sets of experiments on Korea web crawls from June 2006 to July 2006.

First, we generated the in- and out-degree distributions, confirming previous reports on power laws; for instance, the fraction of web pages with i in-links is proportional to $1/i^{2.1}$. We verified these power laws on a more recent crawl. In our second set of experiments we studied the strongly connected components of the Korea Web. Enumerating strongly connected components is the most common way to study the connectivity of web graph. This connectivity analysis is an important part of the research on web graph. This analysis reveals an interesting picture of the Web's macroscopic structure, called bow-tie structure. We show that power law also arises in the distribution of sizes of these connected components. Finally, in our third set of experiments, we studied the evolution of the Korea web graph from the evolving of web pages and link structure by comparing high popularity sites and randomly chosen sites. Then, we analyzed the similarities and differences of the Korea web to the evolution of the global [2] and China Webs [4].

Korea web graph shows many different and similar properties, in both structure and evolution, compared to the global and China web graph. This research on Korea web graph will be useful for predicting the growth scale of Korea Web, improving the performance of Korea web search engine, and processing Korea web information.

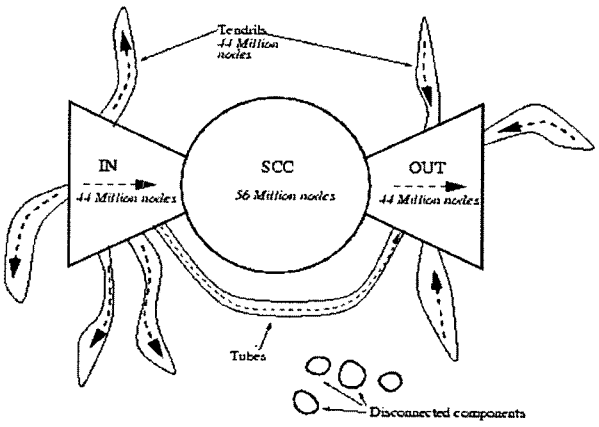
The rest of the paper is organized as follows: In Section 2, we review the previous study work on the global web graph and some regional web graph. Also we introduce algorithm for enumerating SCC (strongly connected components) that we use. In Section 3, we describe our research results from a static snapshot of Korea web graph with several distributions and connectivity of the Korea Web. In Section 4, we present the studies on the evolution of Korea web graph from several points of view. We conclude in Section 5.

2. Related Work

Albert et. al. [1], Broder et. al. [2], and Kumar et. al. [13] studied the degree distribution of nodes in the web graph. They performed empirical studies using graphs of sizes ranging from 325,729 nodes (University of Notre Dame) [1] to 203 million nodes (AltaVista crawler data) [2]. It was found that both the in-degree and out-degree of nodes on the web follow power-law distribution. The number of web pages having a degree i is proportional to $1/i^k$ where $k > 1$. This implies that the probability of finding a node with a large degree is small yet significant. Both [1] and [2] estimated the exponent of in-degree distribution to be 2.1. According to [1], the out-degree distribution has an exponent of 2.45. The empirical study in [2] shows that the exponent for out-degree distribution is 2.72, though the initial segment of the distribution deviates significantly from power law. The study on African web [6] gives further evidence to the power law distribution, the in-degree of African web pages with a exponent of 1.92. G. Liu et. al. [4] reported an in-degree distribution with an exponent 2.05 and out-degree distribution with an exponent 2.62 for the China Web. The average degree of a node in the web has been found to be seven [8].

Broder et. al. [2] also analyzed the structure of the global web graph depicted as a bow-tie. Figure 1 shows this structure. There are four pieces in this structure.

The first region is the giant SCC. There are directed paths from each node in the SCC component to all other nodes in the SCC. There is a set of newly-formed nodes called IN having only outgoing links and another set of introvert nodes called OUT having only incoming links (e.g., some of the corporate and e-commerce sites). There are directed paths from each node in IN to (all nodes in) SCC and directed paths from (all nodes in) SCC to each node in OUT. There is another set of nodes called TENDRILS, which doesn't have any directed paths going to the SCC, nor has any directed paths coming from the SCC. There exists a directed graph from nodes IN to TENDRILS and from TENDRILS to nodes in OUT. Each of IN, OUT, and TENDRILS region occupy about 21% of total number of nodes. Finally, some nodes in TENDRILS from the IN region have edges going to nodes in TENDRILS



(Figure 1) Connectivity of the global web

in the OUT region, forming a TUBE. A small group of remaining nodes are part of disconnected components, which make up about 8% of the Web.

J. Han et. al. [3] showed that the China web graph, manifests different properties from global web graph. The giant, strongly connected component of the China web graph is much bigger than the global web graph.

In the area of web evolution, Ntoulas et. al. [7] performed experiments to study the evolution of the web from a search engine perspective. J. Cho et. al. [5] studied the popularity evolution of web pages. G. Liu et. al. [4] reported on the evolution of China web graph. J. Cho and Garcia-Molina [17] studied the change pattern of 720,000 web pages from 270 web sites selected from various domains (com, edu, gov, net, org) every day for four months. They found that the average change-interval of a web page is about four months (approximately). In their study, more than 70% of the web pages remained unchanged for about one month. It took about 50 days for 50% of the web pages to change or be replaced by a new page. Changes to a web page are random events that can be modeled as a Poisson process this was verified for the web pages in their sample data.

web graph consists of several hundreds of millions of nodes and several billions of edges. Due to this large scale of the web graph, we can hardly load the full graph into the main memory for enumerating SCCs in web graph. To solve this problem, J. Han et. al. [3] proposed an algorithm for enumerating SCCs in web graph by split-merge approach. The basic idea of this algorithm is to split the original graph into parts which are smaller enough to load into the main memory and compute SCCs one by one, and finally merge them together. We used this algorithm for enumerating SCCs of the Korea web for our experiment.

Each step of this algorithm is as follows: First, classify the nodes of the original graph G into n-groups. Build a sub-graph with each group of nodes and the links among them. Next, decompose each sub-graph into SCCs. If the sub-graph is small enough to load into the main memory, use any algorithm for enumerating SCCs. Otherwise, recursively apply the split-merge algorithm. Assume that each SCC in a sub-graph is a node, and eliminate the duplicated links between them. After that, we obtain the contracted graph G', a graph composed of all the SCCs. Decompose the contracted graph G'into SCCs. If the G' is small enough to load into the main memory, use any algorithm for enumerating SCCs. Otherwise, recursively apply the split-merge algorithm. Finally, merge the SCCs from sub-graphs with the help of the decomposition of G'. At last we get SCCs in graph G.

3. Graph Structure of the Korea web

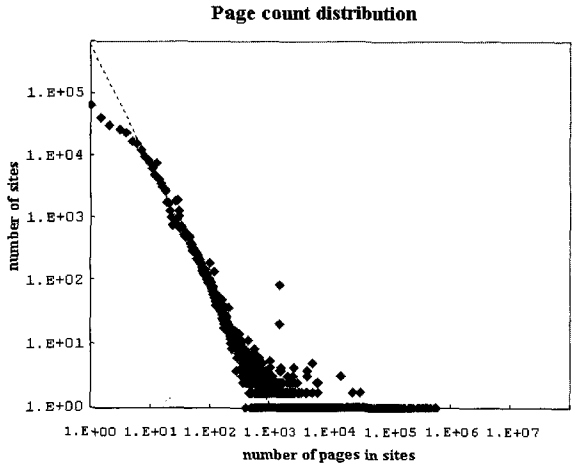
For our experiment, we crawled Korea web pages using our crawler [14] with three machines from June 2006 to July 2006. The IP address is used to determine to see if it belongs to the Korea sites. The crawler downloaded Korea web pages up to the ninth depth, and it collected at most 640,000 web pages from a single site. In order to reduce the size, we compressed links as we downloaded them. The size of the raw data was approximately 65GB, which contain various information on URLs for pages themselves and hyperlinks among pages. We processed the raw data to create Korea web graph, such as removing invalid URLs, and assigning a unique ID to each URL. We finally constructed the Korea web graph, which contains 116 million pages (nodes) and 2.7 billion links (edges).

3.1 Power Law Distributions

We now consider various power law phenomenon of the Korea web graph.

Page Number Distribution in web Sites

Figure 2 shows the distribution of page numbers in web sites. The x-axis shows the number of pages in each site while the y-axis shows the number of sites which have the corresponding x pages. Each point (x, y) on the distribution indicates that y number of sites have x pages. This graph exhibits that the distribution of the number of pages in web sites follows the power law while the exponent is roughly 2.3. Under this distribution, the top 18% of sites possess about 90% of the total pages of the Korea Web, while 82% of sites contain only 10% of the total pages. This implies that the distribution of the number



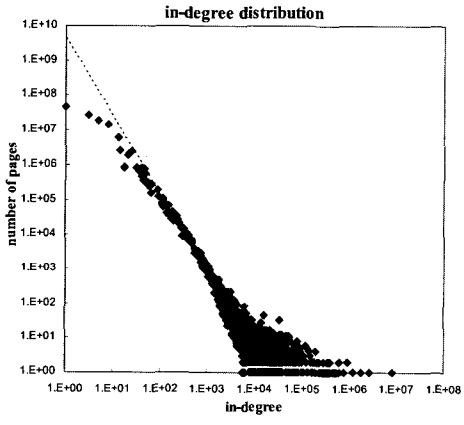
(Figure 2) Page number distribution in web sites

of pages in web sites of the Korea web also obeys Pareto's Law (also known as 20:80 law) although the proportion is a little different. The anomalous points at 1417 and 1418 on the x-axis are due to a cluster of sites that have identical web pages even though they have different host names. The site with most pages has about 590 thousand pages.

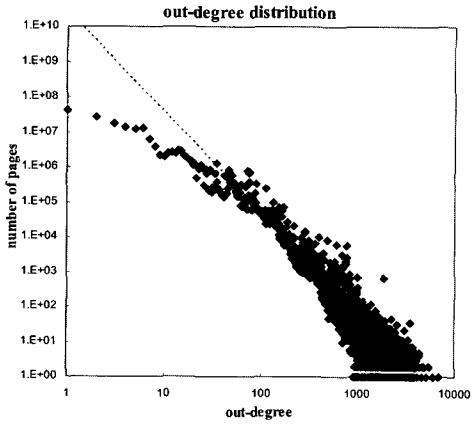
The Korea web is different from the China web which has exponent of 1.74. What it means that the probability that we find web sites in Korea becomes more exponentially decreased than we find such web sites in China, as the number of web pages in a site increases.

Degree Distribution of the Korea web

The degree distributions in the Korea web also follow the power law distribution. Figure 3 and Figure 4 are a log-log plot of the in-degree and out-degree distributions of the Korea web graph, respectively. In all our log-log plots, straight lines are linear regressions for the best power law fit. Figure 3 shows that the distribution of in-degree exhibits a power law with exponent roughly 2.2, which is almost the same value as in the global web [2]



(Figure 3) in-degree distribution



(Figure 4) out-degree distribution

and the China web [4]. Applying an inverse polynomial to the data, we can find the probability that a page has i in-links to be roughly proportional to i-2.2. The page that has the most in-links contains as many as 47 million in-links. The out-degree distribution also exhibits a power law while the exponent is roughly 2.8, as in Figure 4. The average out-links in a page of the Korea web is 27.4. This number is about 3 times more than the average out-links (i.e. eight) which was reported in 1999 [2]. Here we would like to conjecture that the number of links in a page is increasing (at the same time the connectivity among web pages is growing) as time goes by. Note that the pages with out-degrees less than 100 on the x-axis significantly deviate from the best power law fit, suggesting that we might need to have a new distribution to model pages with low out-degrees.

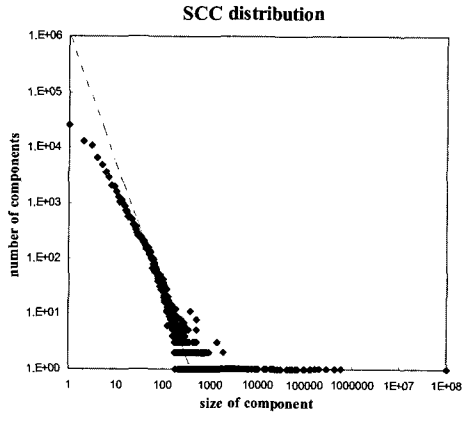
Distribution of Strongly Connected Components

By running the split-merge algorithm [3], we find that there is a single large SCC consisting of about 99 million pages, and all other components are significantly smaller in size. The single large SCC has barely 86% of all the Korean pages. Figure 5 indicates that the distribution of the SCC sizes of the Korea web also follows a power law with exponent roughly 2.3.

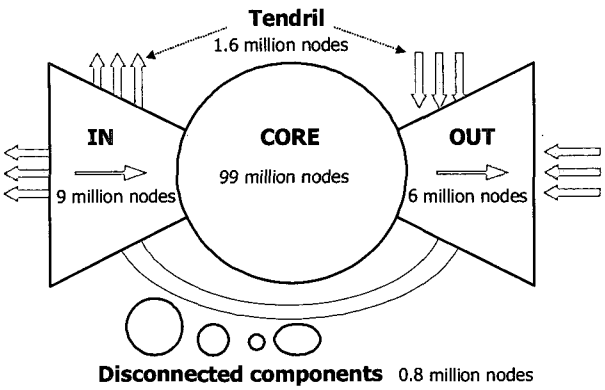
3.2 The Macro Structure of the Korea web

Figure 6 shows the structure of the Korea web graph. The Korea web graph we built contains 116 million pages and 2.7 billion links. The CORE contains about 99 million pages, the IN contains about 9 million pages, the OUT contains about 6 million pages, and the rest contains about 2.4 million pages.

The graph structure of the Korea web exhibits characteristics that are different to the global web graph [2]



(Figure 5) Distribution of strongly connected components



(Figure 6) The bow-tie structure of the Korea web graph

and the China web graph [4]. The CORE possesses around 86% of the pages of the Korea Web. This is higher than the CORE possessed by the 28% in the global web graph [2] and 80% in the China web graph [4]. In other words, the connectivity of the Korea web is higher than the global and China Web. Furthermore, if pages u and v are randomly chosen in the Korea web, the probability that there exists a path each other is at least 0.74 (= 86/100 * 86/100), excluding the existence of many tiny SCCs in the web graph.

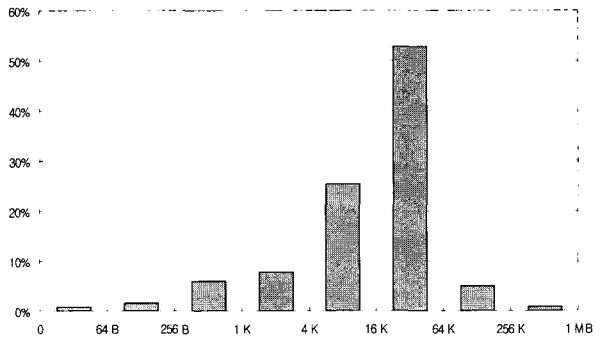
The web service providers are currently interested in providing blog and community services to personal users. Personal users tend to create their web pages under a "personalized" frame, which often automatically put links to famous sites in the newly created pages. Business users also like to put links to famous sites in their web pages for various reasons such as increasing accessibility, advertisement and so on. Above all, fast growing are a few giant portal companies, which provide personal services like communities, blogs, thus allow to create a huge number of new pages. The front-runner companies in Korea would include "www.naver.com", "www.daum.net" and "www.nate.com". As the role of such companies become important, more web users get personalized services (private pages are also made in these sites) and companies connect links to popular sites. After all, the Korea web becomes more and more centralized. The observation leads us to expect the size of CORE continue to increase.

4. Evolution of the Korea web

Recent studies suggest that there are about four billion publicly-indexable web pages in the World-Wide web (simply web). Search engines rely on both the contents and the link structures of web pages to satisfy users' requests (for example, Google uses the PageRank values). The evolution

〈Table 1〉 Domain Distribution

Domain	Fraction of pages in domain
.co.kr	37%
.com	35%
.net	9%
educational (ac.kr, hs.kr, ms.kr, es.kr)	8%
.org or or.kr	7%
.gov or go.kr	1%
Misc.	3%



(Figure 7) Histogram of the sizes of successfully downloaded documents

of the web pages including the link structures is an important and implicate aspect to the search engines. This section describes the statistical analysis on dynamic features of web pages and link modifications of the Korea web. In our experiment, we clustered the web sites into two disjoint sets: popular sites and random sites, and did the same experiments on each set. We also analyze the difference among the Korea, global, and China web in terms of the weekly birth and death rate of pages and links. Our analysis reveals that the Korea web is more dynamic than the global web and China web are.

4.1 Domain and Page size Distributions

Table 1 reports the web page portions in the highest level of domains. The distributions of domains for pages in Korea are different with other webs. The Korea web has the more commercial sites (".com" or ".co.kr"), which are amount to 72%, than the global web (41% in 2004 [7]) has. The Korea web has also more educational sites than the China web (3% in 2005 [4]) has.

Figure 7 shows the distribution of web page sizes that were crawled in our experiment. We found that as much as 53% of all observed pages are between 16 Kbytes and 64 Kbytes in size, which are approximately twice bigger than the study [9] in 2004 where 66.3% of all the ob-

served pages were between 4 Kbytes and 32 Kbytes. It is interesting to note that the study [11] in 1999 reported that 80% of all the observed pages were between 1 Kbyte and 16 Kbytes, which are approximately one fourth of the popular web page sizes in our study. One implication of these studies is that the web page size is definitely growing over time. Our experiment leads us to agree with the argument made by Ntoulas, Cho and Olston [7] that the current growth of the web is mainly driven by the increased size of pages over time, not by the increased number of web pages.

4.2 The Evolution of Pages and Links

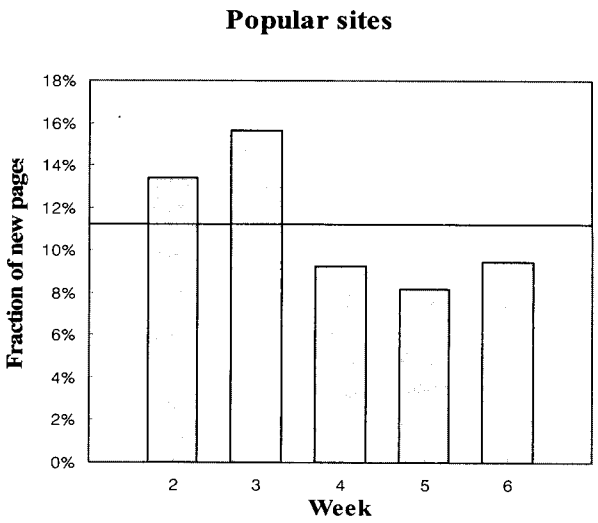
In this section, we investigate the phenomenon, called the evolution of the web that pages and links in the Korea web are dynamically appearing, disappearing and changing with time. This area of the evolution of the Web graph, [4][7] studied about popular sites only. For our evolution study, we used 200 popular sites and 800 randomly chosen sites. We used Google's PageRank as our basis for evaluating popularity [15]. The value of Google PageRank is 0 to 10, the higher value being more important and popular. We decided to think that a site that has PageRank value more than 6 is a popular site. To compare the evolution of average sites with popular sites, we chose 800 sites randomly from various topics.

We collected pages from the sites every week from July 2006 to August 2006, for a total of 6 weeks. While collecting our data, some web sites were temporarily unavailable or our local network connection may have been unreliable. We believe that these glitches were relatively minor in most cases. To counter-balance this short-lived unavailability, our crawler made up to three attempts to download each page.

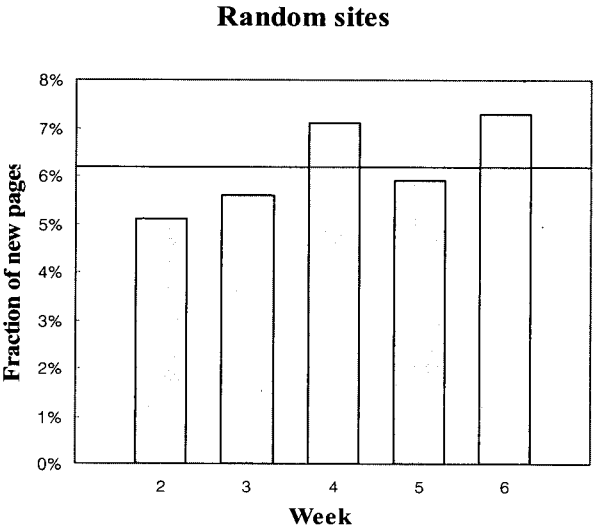
Weekly birth rate of Pages

We first evaluate how many new pages are created every week. For this, we measured what fractions of the pages have not been downloaded before for every week. This fraction represents "the weekly birth rate" of web pages. We use the URLs of pages to distinguish pages. We consider a page to be "new" if we did not discover any page with the same URL before.

Figure 8 and Figure 9 show the weekly birth rate of pages. The y-axis shows the fractions of new pages that we crawled in the given weeks. The line in the middle of the graph denotes the average of all the values, representing the "average weekly birth rate" of the pages. From the two graphs we can observe that the average weekly birth rate is about 11.2% for the popular sites



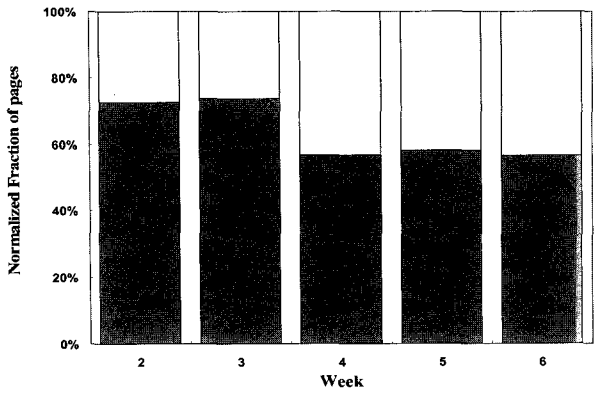
(Figure 8) Fraction of new pages between successive snapshots of popular sites



(Figure 9) Fraction of new pages between successive snapshots of random sites

(Figure 8) and about 6.2% for the random sites (Figure 9). Note that, the average weekly birth rate in the global web (8% in 2004 [7]) and that of the China web (9.8% in 2005 [4]), both for the popular sites. This result shows that new pages weekly created in the Korea web are more than those of the global and China webs for popular sites. It is also worth noting that nearly twice the new pages are created in the popular sites in comparison with to the random sites.

Figure 10 (combines Figure 8 and Figure 9) is a normalized version of the comparison result between popular sites and random sites. The weekly birth rates of popular and random sites are normalized to one. The bottom portion shows fraction of weekly birth rate of popular sites and the top portion corresponds to fraction of weekly



(Figure 10) Normalized fraction of weekly birth rate from popular and random sites

birth rate of random sites. The graph also indicates that the popular sites create more new pages than the random sites do.

4.2.2 Birth and death of Pages

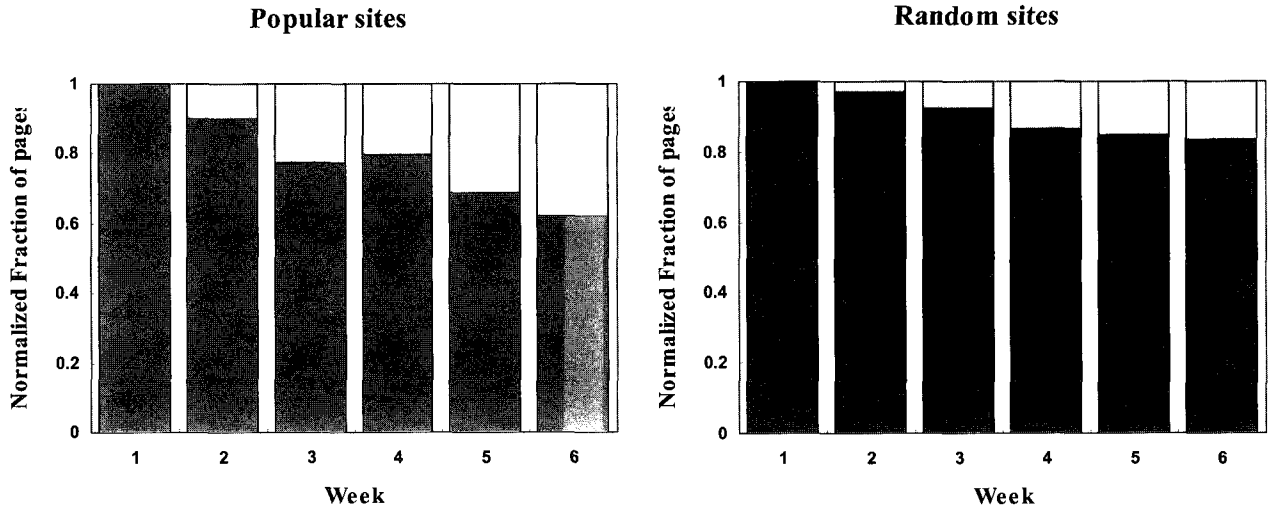
In this experiment, we study how many new pages are created and how many vanish over time. Figure 11 shows the normalized-to-one fractions of old pages (still existing after n weeks since the first crawl) and new pages. The y-axis of Figure 11 denotes the percentages of web pages where the dark bars denote the existing pages after n weeks since the first crawl, and the light bars represent the newly created pages since the first crawl. For example, 90% of the second week pages from the popular sites are old pages from the first week and 10% are newly created pages in the popular sites. We observed that the total number of successfully downloaded pages in each week did not differ significantly. In other words, the number of newly created pages is almost the same as the number of pages disappeared in a given week. We know that six weeks later, only 63% of pages are old against the first week for the popular sites and 83% of pages are old for the random sites. This result shows that the evolution of the popular sites is much faster than the random sites.

Weekly birth rate of Links

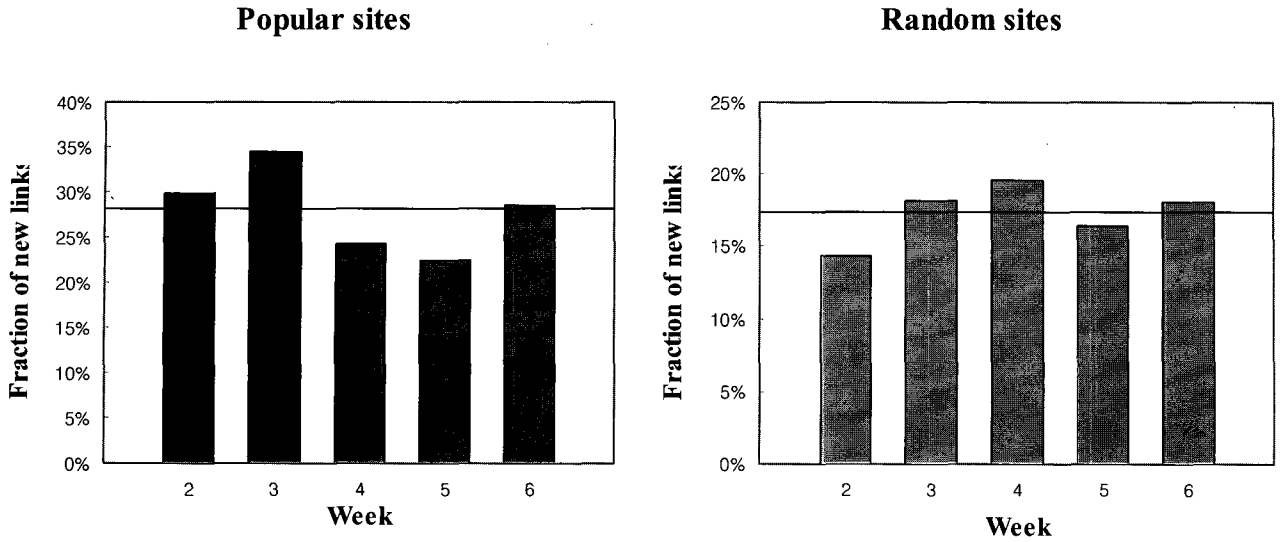
In this experiment, we evaluate how many new links are created every week. In other words, for every week, we measure the fraction of the links that have not been downloaded before. Figure 12 shows the weekly birth rate of links. The y-axis denotes the fraction of new links we crawled in the given week. The line in the middle of the graph represents the average of all the values, representing the "average weekly birth rate" of the links. From the two graphs we can observe that the average weekly birth rate of links is about 28.2% for the popular sites (left graph), and about 17.4% for the random sites (right graph). Recall, the average weekly birth rate of links in the global web (25% in 2004 [7]) and that of the China web (24.7% in 2005 [5]), both for the popular sites. This result indicates that the link structure of the Korea web is more dynamic than the global or China webs for the popular sites. Also, we learn that the structure of links changed more dynamically than that of pages.

Birth and death of Links

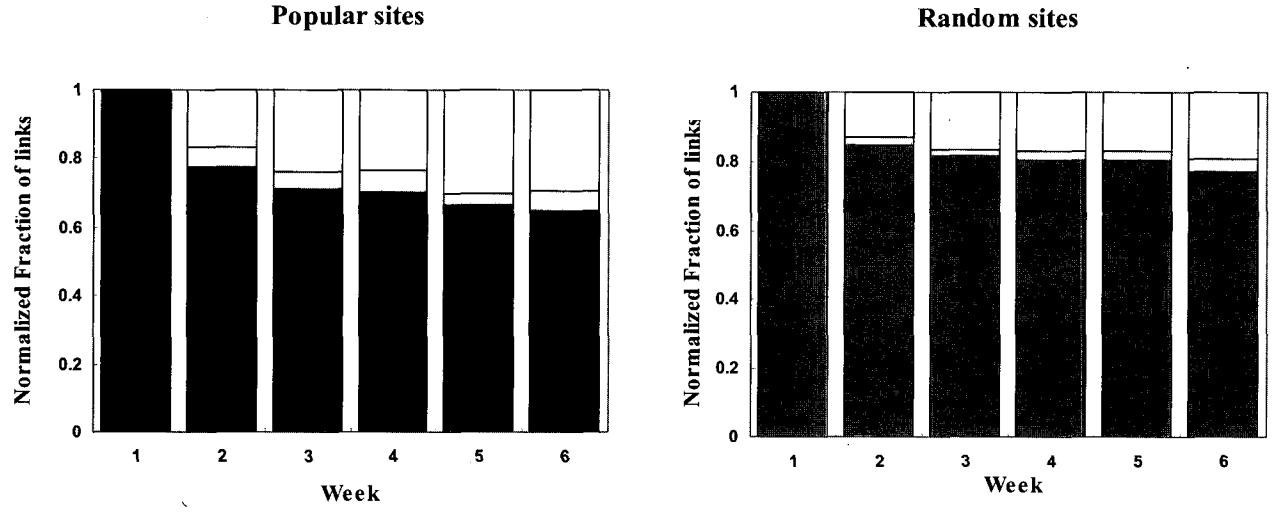
In this experiment, we study how many new links are created from new and old pages and how many disappear over time. The result for this experiment is shown in Figure 13. The y-axis shows the normalized total number



(Figure 11) Normalized fraction of pages from the first crawl still existing after n weeks (dark bars) and new pages (light bars) of popular and random sites



(Figure 12) Fraction of new links between successive snapshots of popular and random sites



(Figure 13) Normalized fraction of links from the first week snapshot still existing after n weeks (dark bars), new links from existing pages (grey bars) and new links from new pages (white bars) of popular and random sites

of links in each snapshot relative to the first week. The dark-bottom portion shows the number of the first-week links that are still present in the given week. The grey portion represents the links that did not exist in the first week, and corresponds to the new links coming from the pages that existed in the first week. The white portion represents the links that did not exist in the first week, and corresponds to the new links coming from new pages.

In these graphs, we can see that most new links are from new pages. From the figure, we know that six weeks later, only 64% of links still remained against the first week for the popular sites, and 77% of links still remained for the random sites. We can also see that the fraction of the new links coming from the new pages of the popular sites is more than the random sites. In other words, hyperlinks are updated more often on the popular sites than on the random sites. This experiment result is similar to the weekly birth rate of links, where the evolution of the popular sites is faster than the random sites. This dynamic change of link structures indicates that search engines may have to update the ranking metric quite often.

5. Conclusion and Future Work

We studied the Korea web graph and analyzed the similarities and differences with respect to the global and China web graphs. According to our analysis, the CORE of the Korea web has the bigger portion than the global web and the China web do. We learn that the Korea web

is highly connected and centralized. We verified power law distributions in the Korea web graph from several aspects, and confirmed, as expected, that power law is a basic property of the Web.

Also We studied the evolution of web pages and link structure, by comparing the popular sites and the randomly chosen sites. We know that the popular sites are dynamic than the random sites. Also, the evolution of the link structure is much faster than that of pages. The change of the link structure is more dynamic than the web pages themselves. The complications of such finding are that system administrators need to often (and also periodically) run web algorithms that use the link structures, in order to precisely model the web.

During the experiment, we found many web communities. A web community is a set of sites that have a similar topic and have many links each other. These communities can be easily found in the form of a single SCC during constructing the web graph. As a perspective of web search engines, such SCC can be utilized to satisfy given topic-oriented user requests, possibly complementing the existing directory services.

The contributions of this paper are three-fold: First, we construct the Korea web graph and analyze them. Also we compared the Korea web graph with other webs. Second, we investigate the random sites too and learned that the popular sites are over twice dynamic than the random sites. Third, we also found that the Korea web changes more dynamically than other parts of world (in particular in China) do. One possible reason about this phenomenon would be that the Korea web has proportionally more commercial sites than other webs have. As for future work, we plan to study how much the change of link structure effects in terms of page ranking algorithms.

References

[1] R. Albert, H. Jeong and A. - L. Barabasi, "Diameter of the world wide web," Nature, 401(6749), 1999.

[2] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener, "Graph structure in the web," the 9th International World-Wide web Conference, 2000.

[3] J. Han, Y. Yu, G. Liu and G. Xue, "An Algorithm for Enumerating SCCs in web Graph," the 7th Asia Pacific web Conference, pp.655-667, 2005.

[4] G. Liu, Y. Yu, J. Han and G. Xue, "China web Graph Measurements and Evolutions," the 7th Asia Pacific web Conference, pp.668-679, 2005.

[5] J. Cho and S. Roy, "Impact of search engines on page popularity," the 13th World-Wide web Conference, 2004

[6] P. Boldi, B. Codenotti, M. Santini and S. Vigna, "Structural properties of the African web," 2002.

[7] A. Ntoulas, J. Cho, and C. Olston, "What's New on the Web? The Evolution of the web from a Search Engine Perspective," In Proceedings of the 13th International World Wide web Conference, pp.1-12, 2004.

[8] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins and E. Upfal, "The web as a graph," Lecture Notes in Computer Science, 1627, 1999.

[9] D. Fetterly, M. Manasse, M. Najork, and J. L. Wiener, "A Large-Scale Study of the Evolution of web Pages," In Software: Practice and Experience, Vol.34, No.2, pp.213-237, 2004.

[10] Z. Bar-Yossef, A. Berg, S. Chien, J. Fakcharoenphol and D. Weitz, "Approximating aggregate queries about web pages via random walks," the 26th VLDB Conference, 2000.

[11] A. Heydon and M. Najork, "Mercator: A scalable, extensible web crawler," the 8th International World-Wide web Conference, pp.219-229, 1999.

[12] A. Barabasi and R. Albert, "Emergence of scaling in random networks," Science, pp.509-512, 1999.

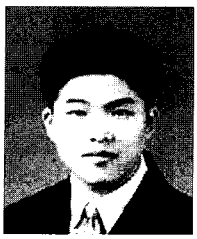
[13] S. R. Kumar, P. Raghavan, S. Rajagopalan and A. Tomkins, "Trawling emerging cyber-communities automatically," the 8th World-Wide web Conference, 1999.

[14] S. J. Kim, S. H., Lee, and H. J. Kim, "Implementation of web Robot and Statistics on the Korean Web," In Proceedings of the 2nd International Conference on Human.Society@ Internet, pp.341-350, 2003.

[15] Google PageRank. http://www.google.co.kr

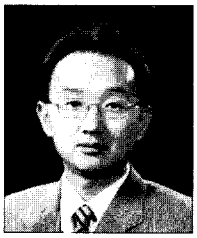
[16] B. Hayes. Graph theory in practice: part I. American Scientist, 88(1):9-13, Jan. 2000.

[17] Cho, J., Garcia-Molina, H.: The evolution of the web and implications for an incremental crawler. Stanford University, CA, (1999).



한 인 규

e-mail : ikhan@comp.ssu.ac.kr
2004년 세종대학교 컴퓨터학부 (학사)
2007년 숭실대학교 대학원 컴퓨터학과 (석사)
관심분야 : 데이터베이스, 인터넷 데이터베이스



이 상 호

e-mail : shlee@comp.ssu.ac.kr
1984년 서울대학교 컴퓨터공학과 (학사)
1986년 미국 노스웨스턴 대학교 전산학과 (석사)
1989년 미국 노스웨스턴 대학교 전산학과 (박사)
1990년~1992년 한국전자통신연구원 선임연구원
1999년~2000년 미국 조지메이슨 대학교 소프트웨어정보공학과 교환 교수
1992년~현 재 숭실대학교 컴퓨터학과 교수
관심분야 : 인터넷 데이터베이스, 데이터베이스 튜닝 및 성능평가