Chapter 5

Experiments

5.1 Experimental Dataset

The data used in our experiment is a large-scale crawling of Japanese web sites collected in May 2004, including 5.8 million of nodes and 283 million of edges. The format of nodes (sites) has three tiers (i.e. http://A/B/C). Crawlers stopped gathering the sites when they could not find any Japanese sites in the sub-site of the site.

Many previous papers have observed that various properties of the web graph follow a Zipfian distribution (a function of the form $1/n^k$). We also did some examinations to see the fraction of web sites of some properties in our data file.

Figure 5.1, 5,2 and 5.3 shows the attributes of our data file in edge weight (the number of links between two web sites), indegree (the number of inlinks of a web site) and outdegree (the number of outlinks of a web site).

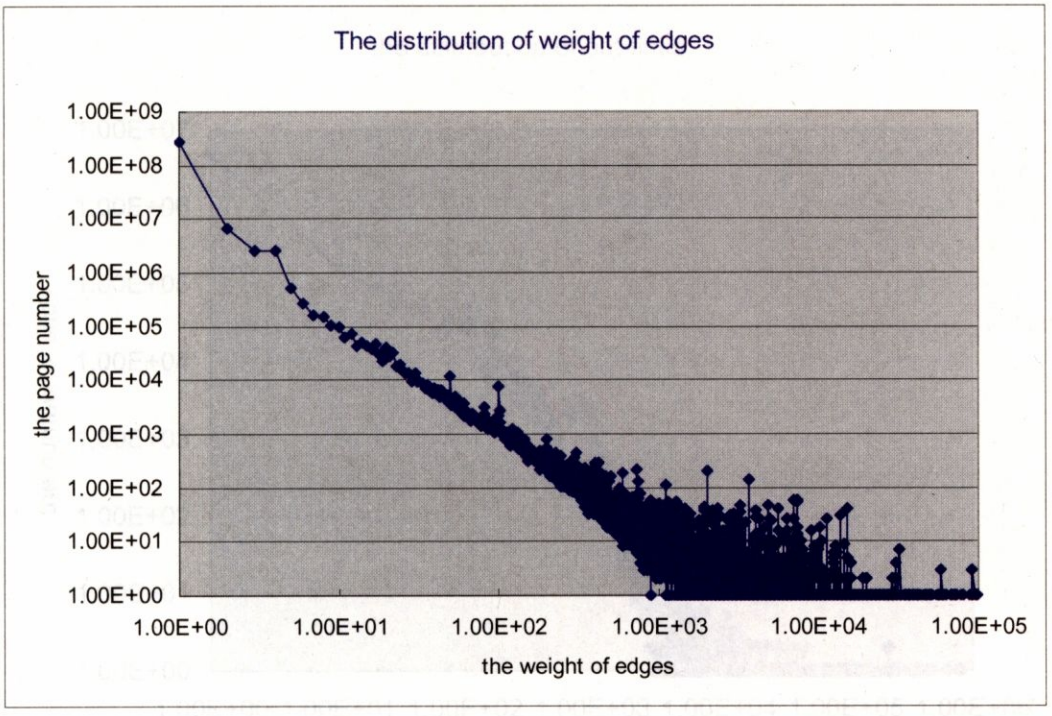


Figure 5.1 Edge weights are Zipf distributed

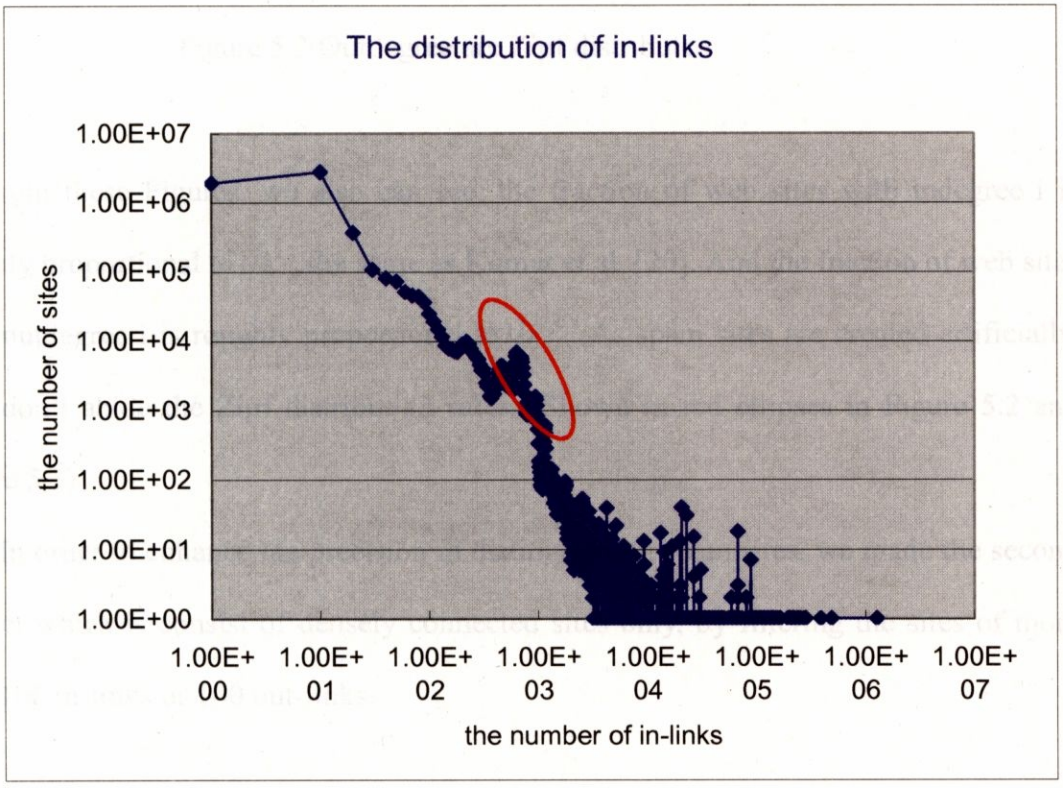


Figure 5.2 Indegree are Zipf distributed

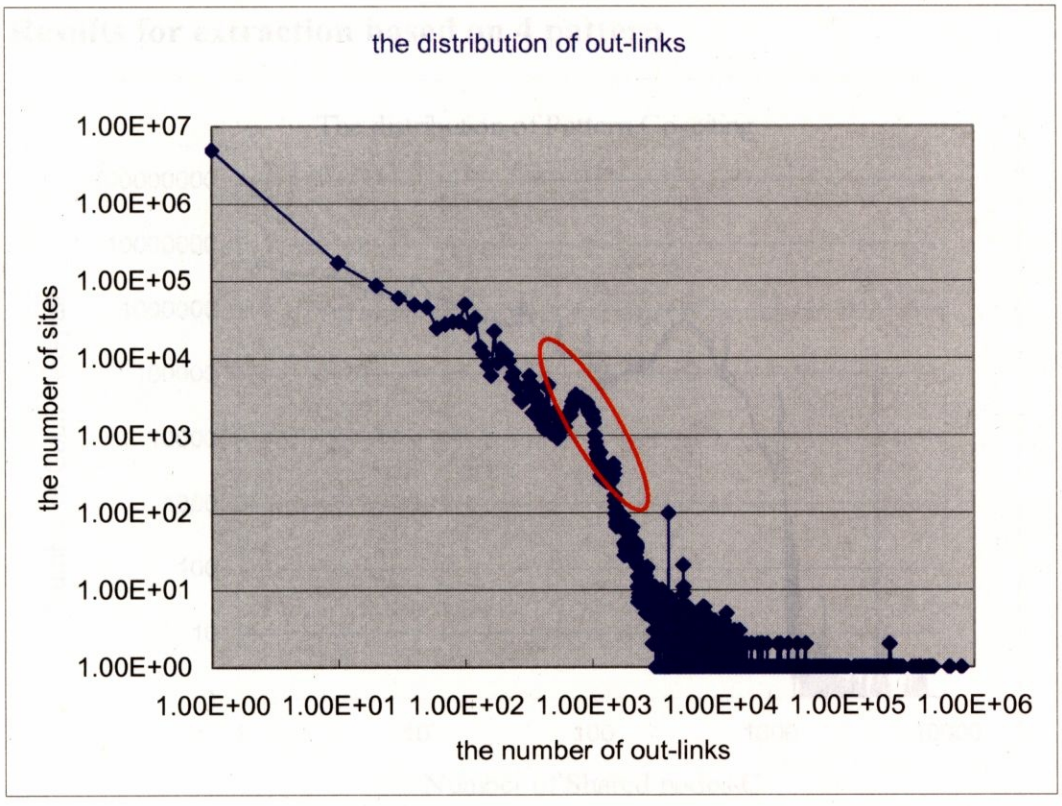
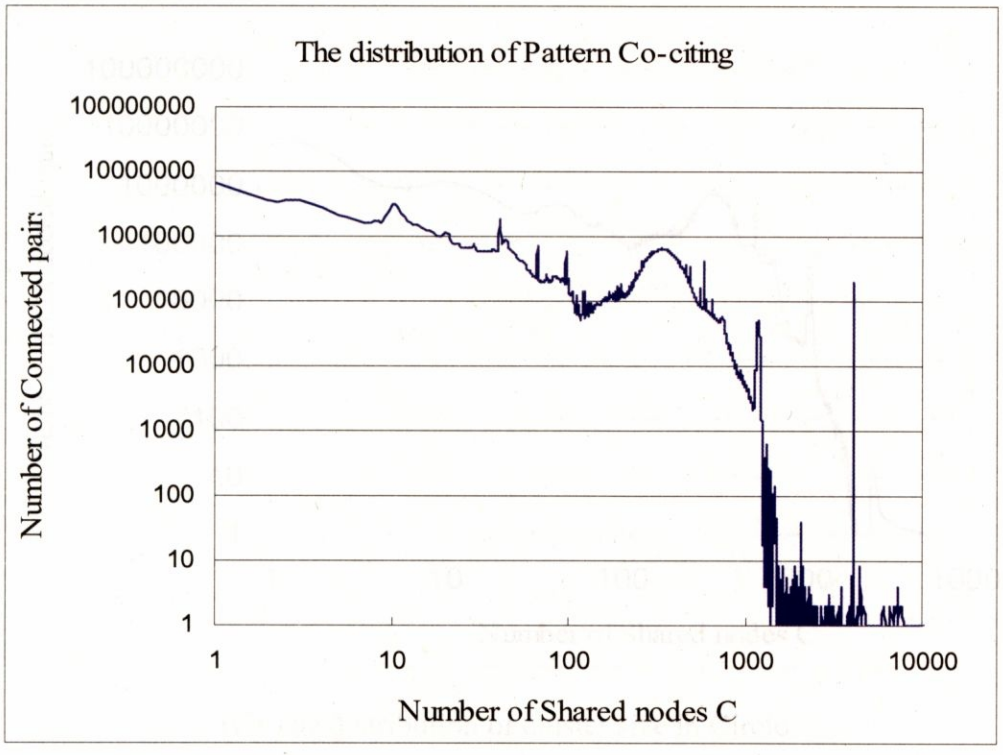


Figure 5.3 Outdegree are Zipf distributed

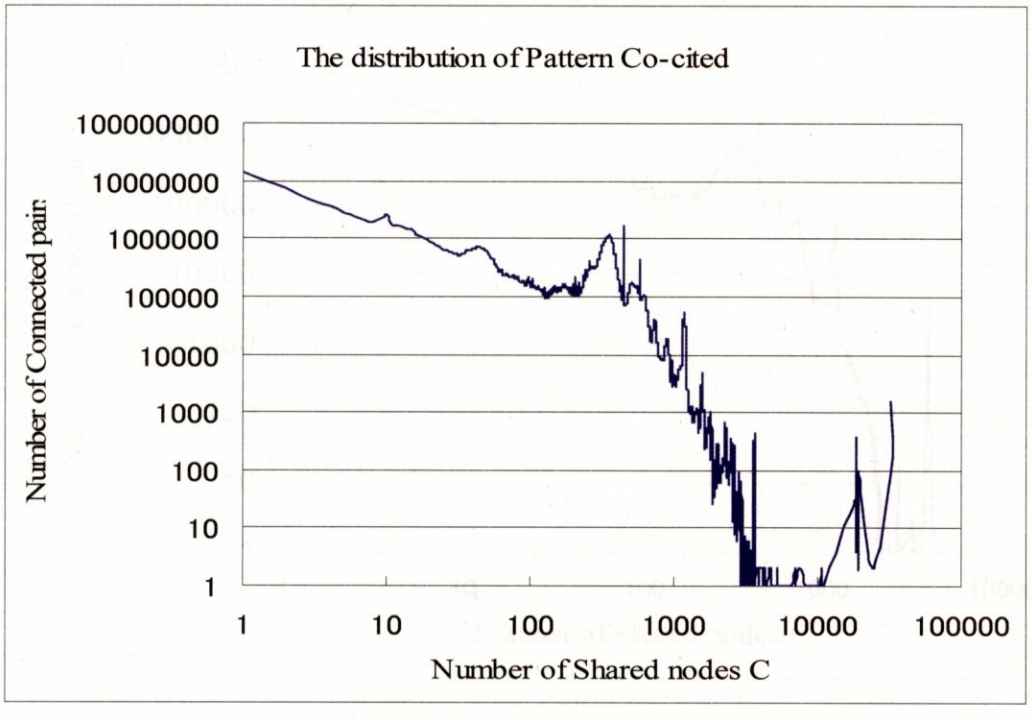
From these Figures, we also can see, the fraction of web sites with indegree i is roughly proportional to $1/i^2$, the same as Kumar et al. [20]. And the fraction of web sites with outdegree i is roughly proportional to $1/i^{2.4}$. As spam sites are created artificially, they don't abide the Zipf distribution rules. (Shown in red ellipses in Figure 5.2 and Figure 5.3)

In order to enhance the precision in distinguishing spam sites, we made the second dataset which is consist of densely connected sites only, by filtering the sites of more than 100 in-links or 100 out-links.

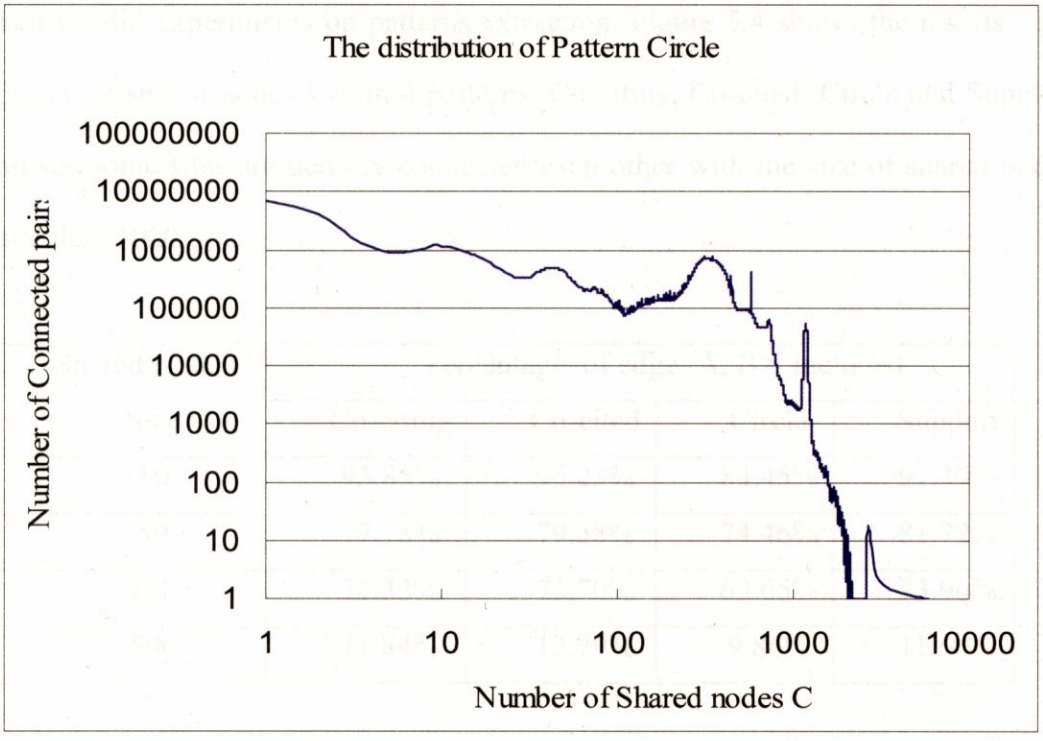
5.2 Results for extraction based on 4 pattern



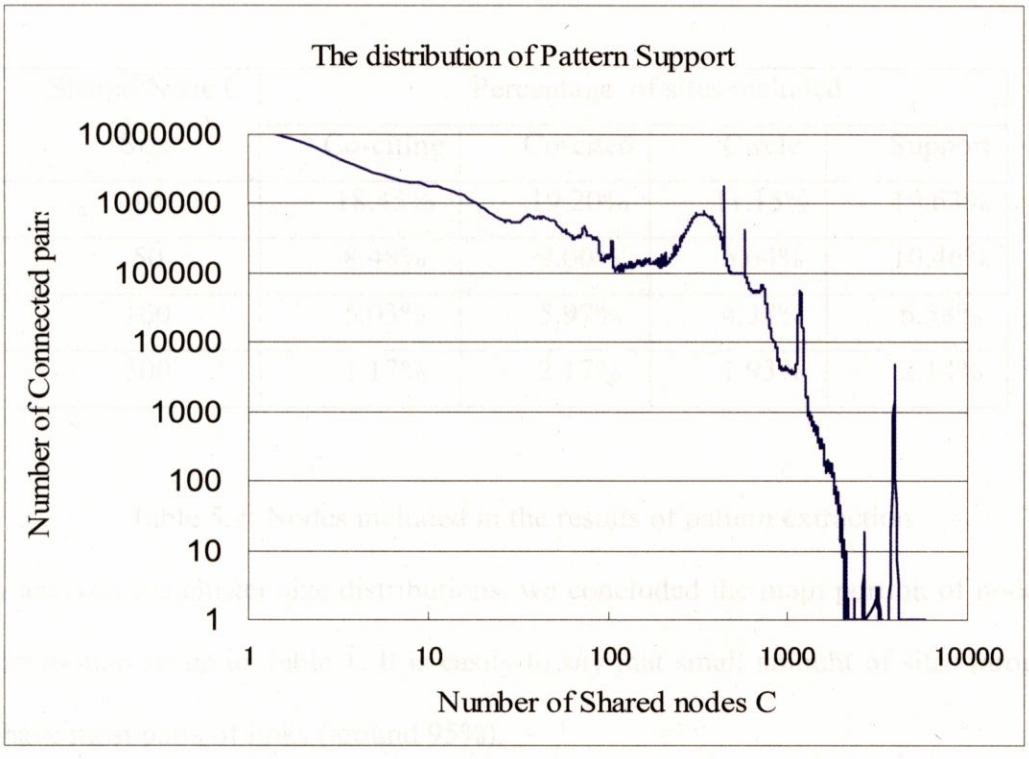
(A) The distribution of cluster size in Co-citing



(B) The distribution of cluster size in Co-cited



(C) The distribution of cluster size in Circle



(D) The distribution of cluster size in Support

Figure 5.4: The distribution of shared nodes size in 4 patterns

Then we did experiments on patterns extraction. Figure 5.4 shows the results—the distribution of shared nodes size in 4 patterns: Co-citing, Co-cited, Circle and Support. We can see some sites are densely connected each other with the size of shared nodes are more than 1000.

Shared Node C Size	Percentage of edge<A, B> included			
	Co-citing	Co-cited	Circle	Support
10	95.85%	95.23%	84.45%	97.39%
50	77.78%	79.58%	74.46%	81.72%
100	71.33%	73.70%	69.65%	73.96%
500	11.84%	12.79%	9.86%	11.60%

Table 5.1: Edges included in the results of pattern extraction

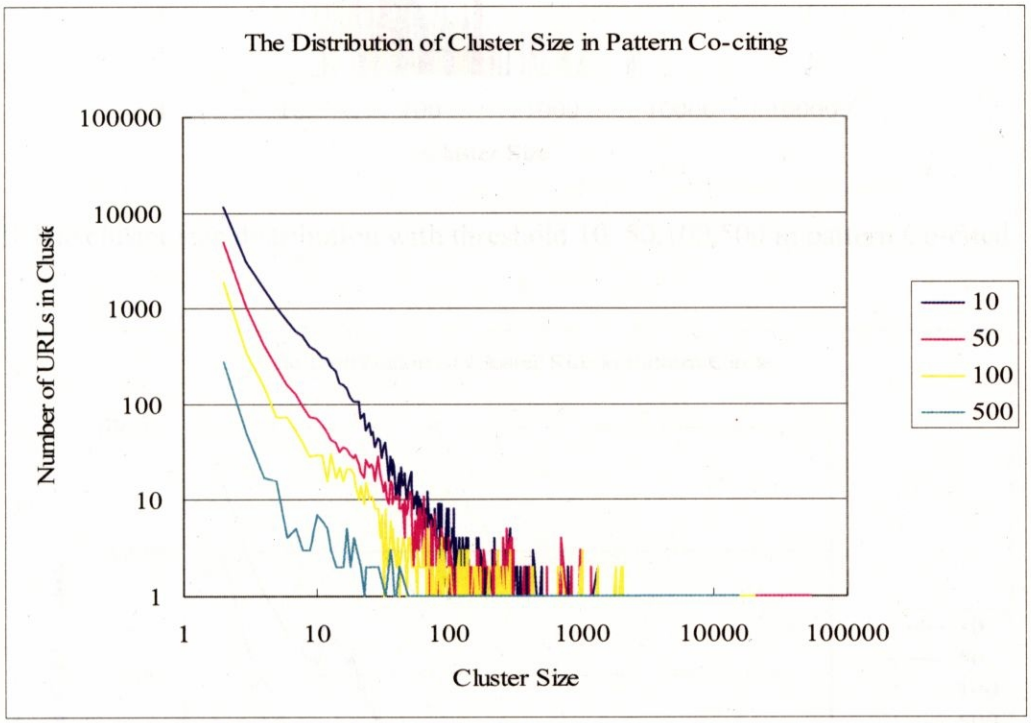
Shared Node C Size	Percentage of sites included			
	Co-citing	Co-cited	Circle	Support
10	18.45%	19.20%	11.15%	19.63%
50	8.48%	9.00%	6.64%	10.46%
100	5.03%	5.97%	4.38%	6.58%
500	1.17%	2.17%	1.93%	2.14%

Table 5.2: Nodes included in the results of pattern extraction

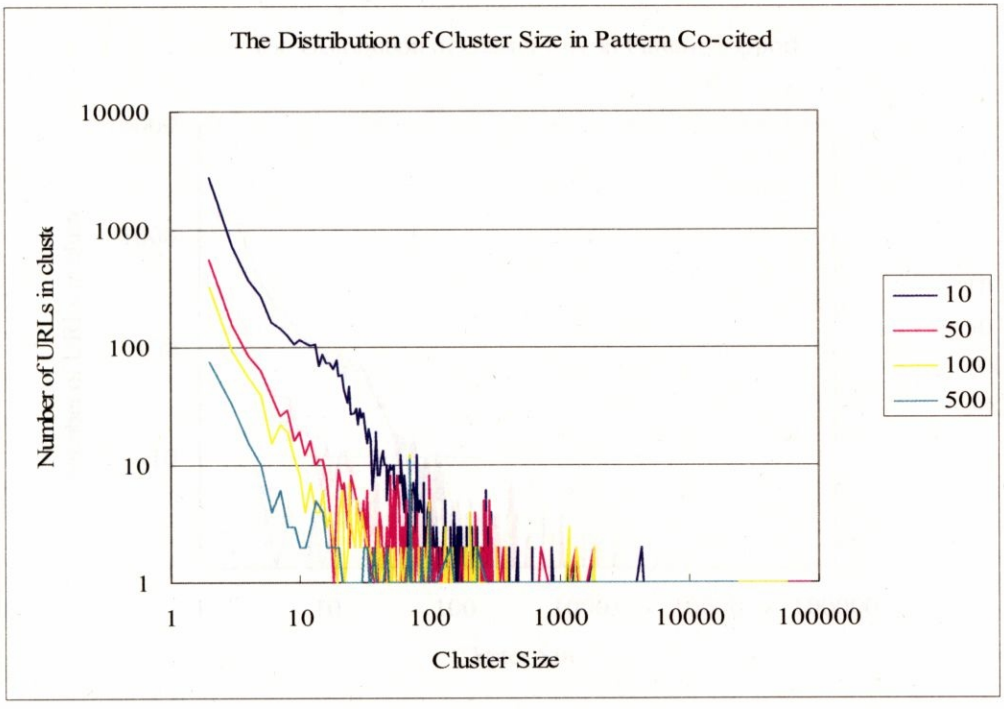
By analysis the cluster size distributions, we concluded the main portion of node C size distribution range in Table 1. It is easily to say that small amount of sites (around 20%) have most parts of links (around 95%).

5.3 Results for Union-find cluster

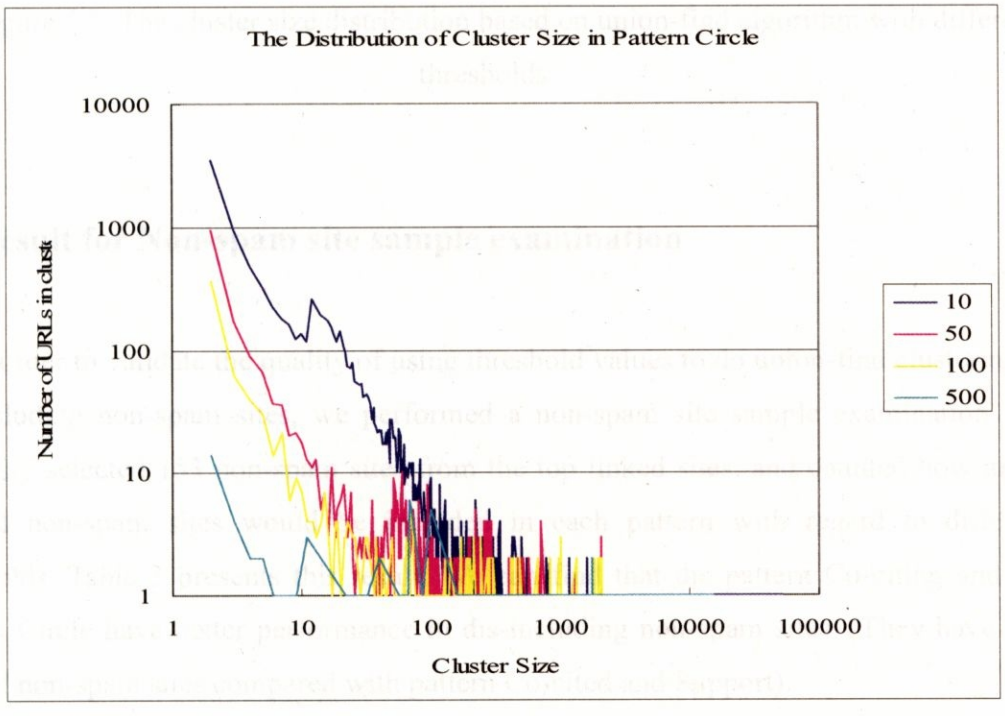
Based on the previous step, we set the thresholds of node C size for merging nodes A and B as 10, 50, 100, and 500 in union-find based clustering, to see how cluster size distribution changes with different threshold. Figure 5.5 shows the experimental result of the cluster size distribution based on union-find algorithms with different thresholds.



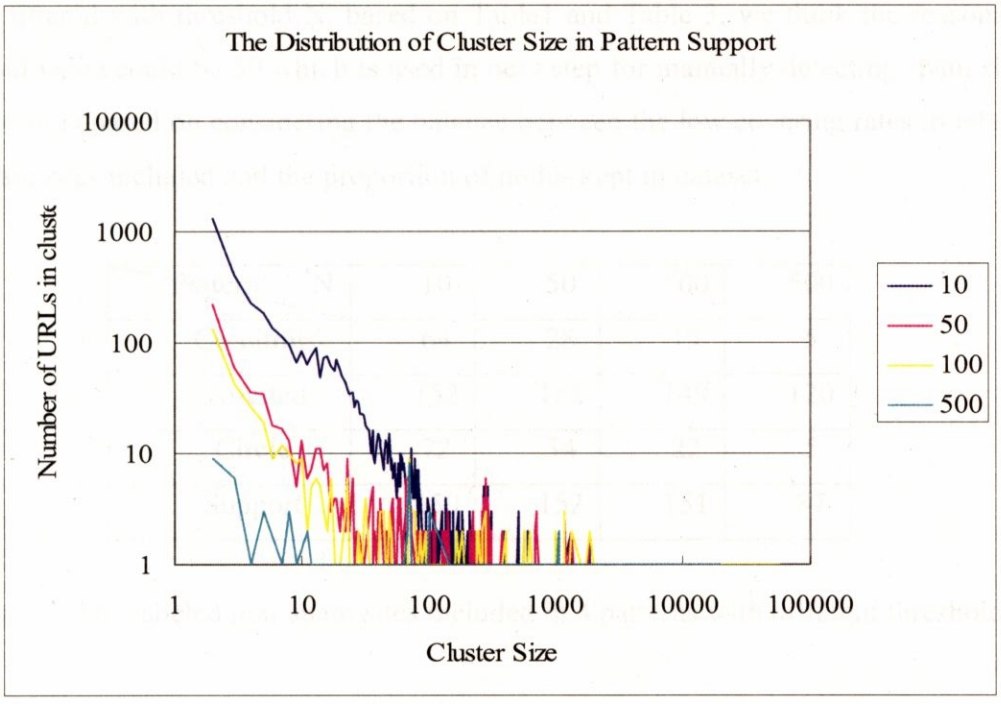
(A) The cluster size distribution with threshold 10, 50,100,500 in pattern Co-citing



(B) The cluster size distribution with threshold 10, 50,100,500 in pattern Co-cited



(C) The cluster size distribution with threshold 10, 50,100,500 in pattern Circle



(D) The cluster size distribution with threshold 10, 50,100,500 in pattern Support

Figure 5.5: The cluster size distribution based on union-find algorithm with different thresholds

5.4 Result for Non-spam site sample examination

In order to validate the quality of using threshold values to do union-find clustering in dis-including non-spam sites, we performed a non-spam site sample examination. We manually selected 153 non-spam sites from the top linked sites, and counted how many labeled non-spam sites would be included in each pattern with regard to different thresholds. Table 3 presents this result. We can find that the pattern Co-citing and the pattern Circle have better performance in dis-including non-spam sites. (They have less labeled non-spam sites compared with pattern Co-cited and Support).

Combining our analysis on the results of the proportion of nodes remaining after

having filtered with threshold N, based on Table1 and Table 3, we think the reasonable threshold value could be 50 which is used in next step for manually detecting spam sites. The reason is based on considering the balance between the low covering rates in labeled non-spam sites included and the proportion of nodes kept in dataset.

Pattern \ N	10	50	100	500
Co-citing	64	28	19	5
Co-cited	152	152	149	120
Circle	72	34	22	5
Support	150	152	151	87

Table 5.3: The labeled non-spam sites included in 4 patterns with different thresholds

As we are curious to know which kind of labeled non-spam sites are included in our experiment on link spam detection, and they are in the same cluster which might have an amazingly huge size or they are in the different clusters. Therefore, we extracted all the URL of labeled non-spam sites included in pattern Co-citing and pattern Circle.

Table 5.4 reveals all the URLs of labeled non-spam sites that are included in the results of Co-citing. By checking the size of clusters that include white sites, we found in pattern Co-citing, most of them are all in the one big cluster which size is 52267. With highly accuracy, in co-citing pattern, there is one big bipartite core that includes many famous sites. This would help to understand, famous site-high visit rates sites-are sharing the benefit from gaining Pagerank score by interconnect 52267 is the biggest size of all the cluster sizes. Therefore, dis-include this cluster would enhance the accuracy of spam detection.

Table 5.5 presents the labeled non-spam URLs included in Pattern Circle. We can see that, these labeled non-spam sites are dispersed distribution, and it's difficult to dis-include them just by link structure based extraction.

Number of In-links	Url	Cluster Size
9910	headlines.yahoo.co.jp/	52267
3285	kids.yahoo.co.jp/	52267
7026	www.2ch.net/	52267
4062	www.aaacafe.ne.jp/	52267
4227	www.asahi.com/	52267
3256	www.biglobe.ne.jp/	52267
3966	www.excite.co.jp/	52267
10261	www.forest.impress.co.jp/	52267
6528	www.fresheye.com/	52267
5527	www.geocities.jp/	52267
21212	www.google.co.jp/	52267
61006	www.google.com/	52267
5434	www.google.com/intl/ja/	52267
4913	www.hatena.ne.jp/	52267
17562	www.infoseek.co.jp/	52267
3214	www.jikoku.com/	52267
3532	www.kakaku.com/	52267
4328	www.kyodo.co.jp/	52267
5155	www.maff.go.jp/	52267
7491	www.mapfan.com/	52267
5865	www.melma.com/	52267
3305	www.metro.tokyo.jp/	2
6775	www.mlit.go.jp/	52267
3876	www.msn.com/	2
5114	www.nifty.com/	52267
4697	www.tinami.com/	52267
3990	www.toshiba.co.jp/	52267
3558	www2.odn.ne.jp/	52267

Table 5.4: 28 labeled non-spam sites included in Co-citing with threshold 50

Number of In-links	URL	Cluster Size
9910	headlines.yahoo.co.jp/	7
3285	kids.yahoo.co.jp/	1792
7026	www.2ch.net/	17
4656	www.alibaba.com/	26
5877	www.ana.co.jp/	9
12907	www.asahi.com/	5
3256	www.biglobe.ne.jp/	262
3386	www.books.or.jp/	2
4431	www.env.go.jp/	92
3966	www.excite.co.jp/	1792
10261	www.forest.impress.co.jp/	2
6528	www.fresheye.com/	1792
5527	www.geocities.jp/	1792
21212	www.google.co.jp/	1792
61006	www.google.com/	1792
4913	www.hatena.ne.jp/	1792
3634	www.hitachi.co.jp/	2
17562	www.infoseek.co.jp/	1792
5155	www.maff.go.jp/	95
16158	www.mag2.com/	1792
7491	www.mapfan.com/	1792
5865	www.melma.com/	1792
5379	www.meti.go.jp/	95
3305	www.metro.tokyo.jp/	5
6164	www.mext.go.jp/	95
9502	www.mhlw.go.jp/	95
6775	www.mlit.go.jp/	95
4064	www.mof.go.jp/	95
20550	www.movabletype.org/	2

4030	www.soumu.go.jp/	95
4697	www.tinami.com/	1792
2971	www.toyoko-inn.com/	1792
47112	www.yahoo.co.jp/	7
3558	www2.odn.ne.jp/	1792

Table 5.5: 34 labeled non-spam sites included in Circle with threshold 50

5.5 Results for spam detection

Even though we concluded pattern Co-citing and Circle with merging threshold value 50 are appropriate for dis-including non-spam web sites, to ascertain which pattern and which threshold are suitable for link spam detection. The main remained work now is to manually check the content of sites in our clustering results in all patterns with different thresholds and educe the statistic sample distribution results in tables.

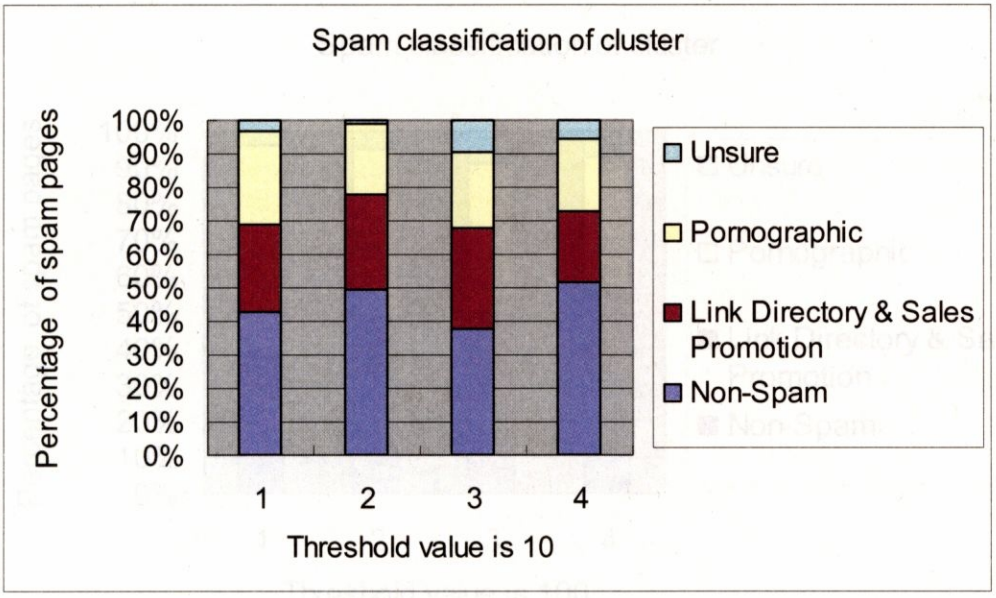
The implementation detail is described as follows:

As the sites' ID is based on their URL, i.e. alphabet order, for redress the balance, we choose the cluster by average distance in the result order.

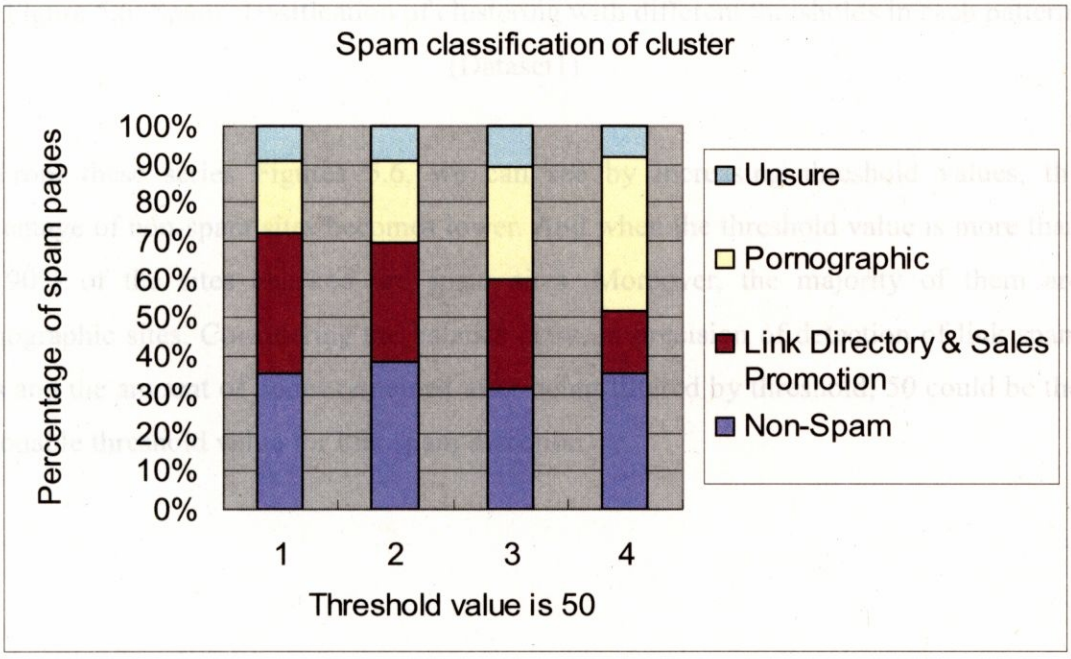
For each pattern, with each threshold 10, 50, and 100, we chose 100 cluster, and checked the first site in each cluster.

We finally compared the results of 4 patterns.

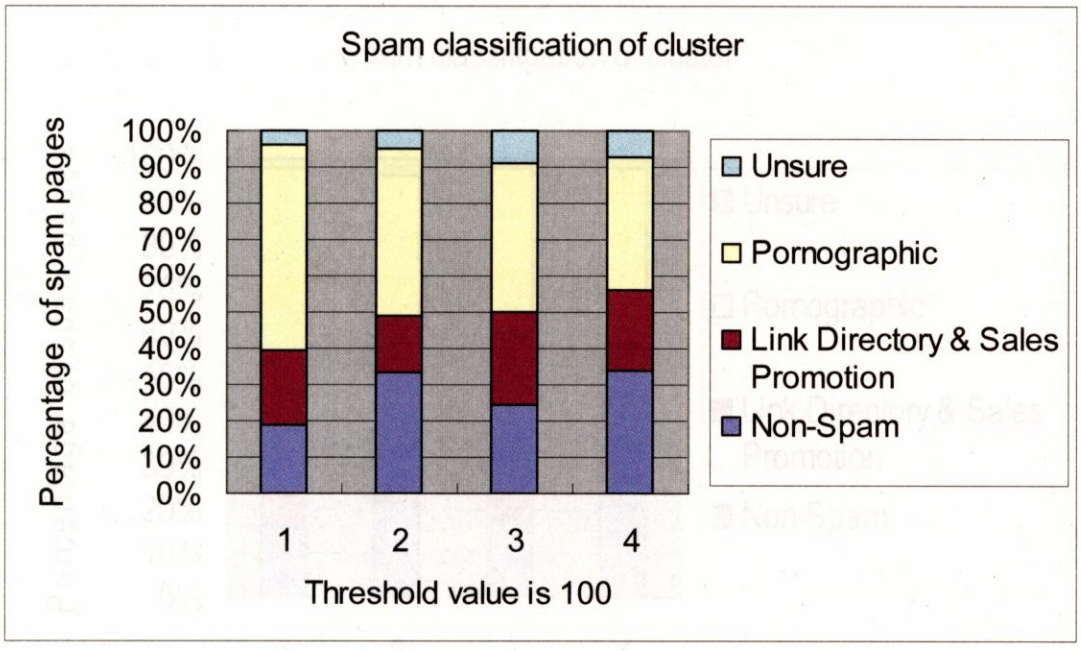
Then, we did the same experiment in the Dataset 2 with the destination of how the precision and coverage rate will change if the site graph dataset become dense.



Threshold 10



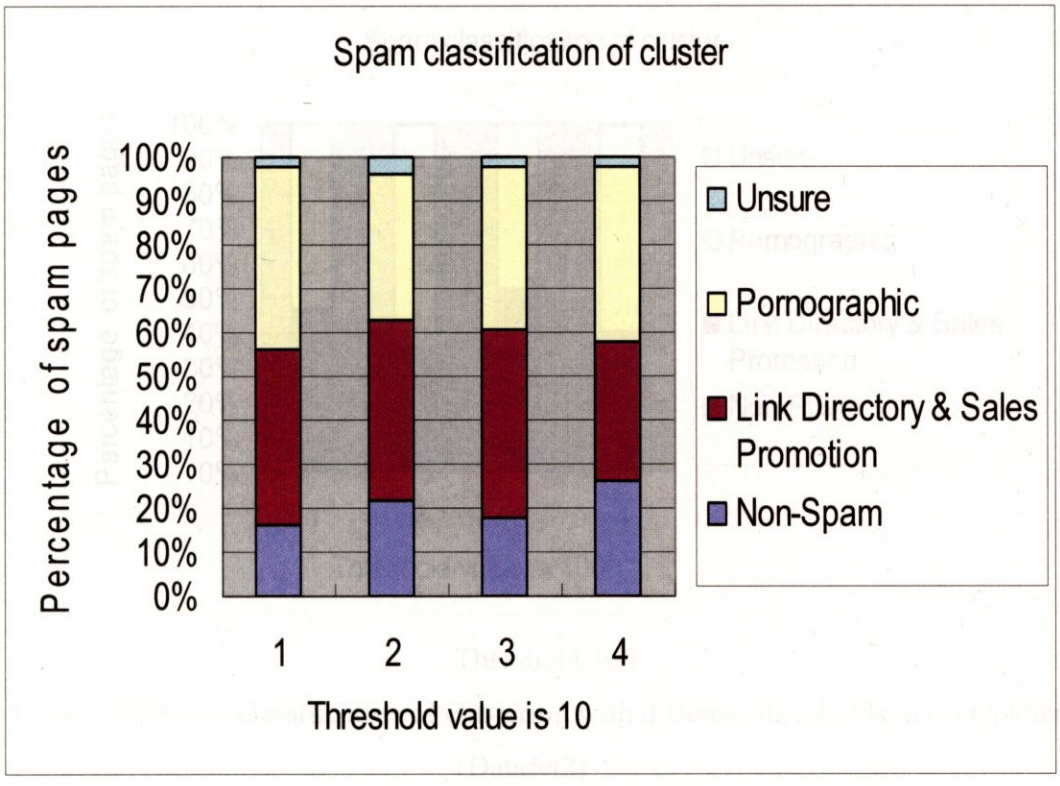
Threshold 50



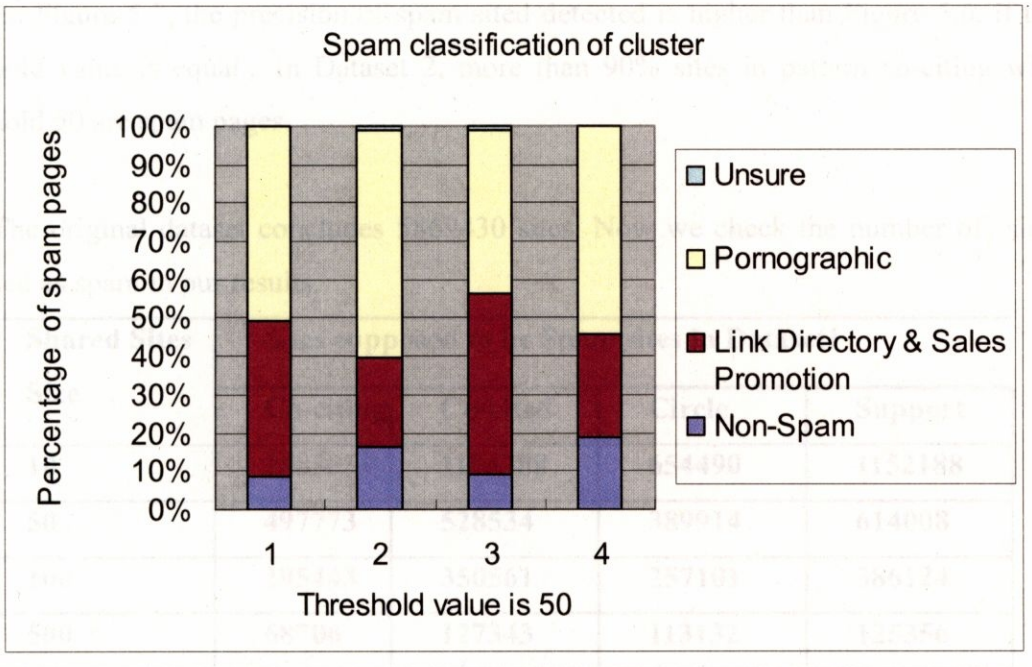
Threshold 100

Figure 5.6: Spam classification of clustering with different thresholds in each pattern (Dataset1)

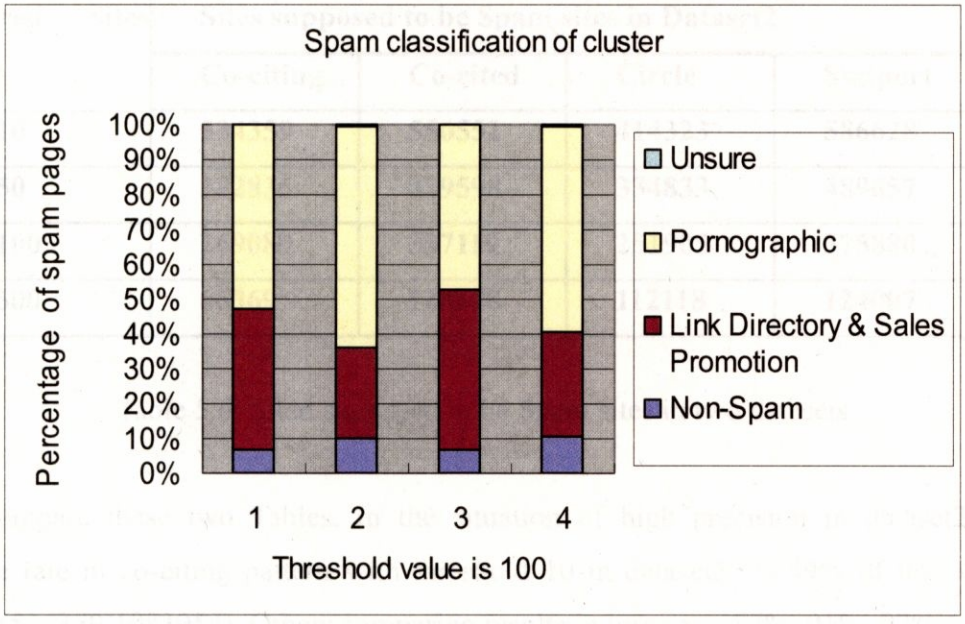
From these series Figures 5.6, we can see by increasing threshold values, the percentage of non-spam sites becomes lower. And when the threshold value is more than 50, 90% of the sites checked are spam sites. Moreover, the majority of them are pornographic sites. Considering the balance between precision of detection of link spam sites and the amount of nodes remained after being filtered by threshold, 50 could be the reasonable threshold value for link spam detection.



Threshold 10



Threshold 50



Threshold 100

Figure 5.7: Spam classification of clustering with different thresholds in each pattern (Dataset2)

In Figure 5.7, the precision of spam sited detected is higher than Figure 5.6, if the threshold value is equal. In Dataset 2, more than 90% sites in pattern co-citing with threshold 50 are spam pages.

The original dataset concludes 5869430 sites. Now we check the number of sites detected as spam in our results.

Shared Sites Size	Sites supposed to be Spam sites in Dataset1			
	Co-citing	Co-cited	Circle	Support
10	1083054	1126788	654490	1152188
50	497773	528534	389914	614008
100	295443	350561	257101	386124
500	68706	127343	113132	125356

Shared Sites Size	Sites supposed to be Spam sites in Dataset2			
	Co-citing	Co-cited	Circle	Support
10	534359	556552	414323	586628
50	382835	229598	334833	489657
100	269089	337119	251905	375880
500	66369	124858	112118	124007

Table 5.6: Sited supposed to be Spam sites in two datasets.

Compare these two Tables, in the situation of high precision in dataset2, the coverage rate in co-citing pattern with threshold 10 in dataset2 is 49% of the one in dataset1(534359/1083054). Others comparing results in turns are: 77%, 91%, 97%.

5.6 Results of combination of pattern 1 and pattern 3(Dataset2)

Considering all the results in the previous steps, we can conclude pattern Co-citing and Circle have good performance in link spam detection; therefore, we combined these two and repeated the same experiments in non-spam site test.

Threshold	10	50	100
No. of non-spam	27	10	3

Table 5.7: The labeled non-spam sites included with different threshold

ID	URLs	Cluster size
1488954	kids.yahoo.co.jp/	13
2788708	www.2ch.net/	12
3831377	www.google.co.jp/	13
3831510	www.google.com/	1005
3897201	www.hatena.ne.jp/	1005

4041297	www.infoseek.co.jp/	1005
4362032	www.maff.go.jp/	12
4377266	www.mapfan.com/	1005
4422007	www.melma.com/	1005
5671923	www2.odn.ne.jp/	1005

Table 5.8:URLs of Labeled non-spam sites included in Co-citing with threshold 50

Table 5.7 and Table 5.8 show the non-spam sites test. Comparing with Table 5.3, it really has a great enhancement in dis-including labeled non-spam sites.

5.7 Visualization of spam sites in cluster units

In order to have an intuitionistic understanding of link spam structures, we make some visualization of spam sites in cluster units (See Chapter 4.5).

Figure 5.8 shows the link structure of one cluster with the size of 3. The position of target sites (presented by ellipses) are easy to confirm and the contents of target sites are on pornography. URLs are presented in Table 5.8. (Shared nodes are also shown).

Figure 5.9 shows the typical link structure of two target sites which main contents are link directory. (Shared nodes are also shown).

Figure 5.10 shows the neighbor hood link structure of one cluster which is composed of 29 sites, where most of these sites are about sales promotions. (Only connected pairs nodes)

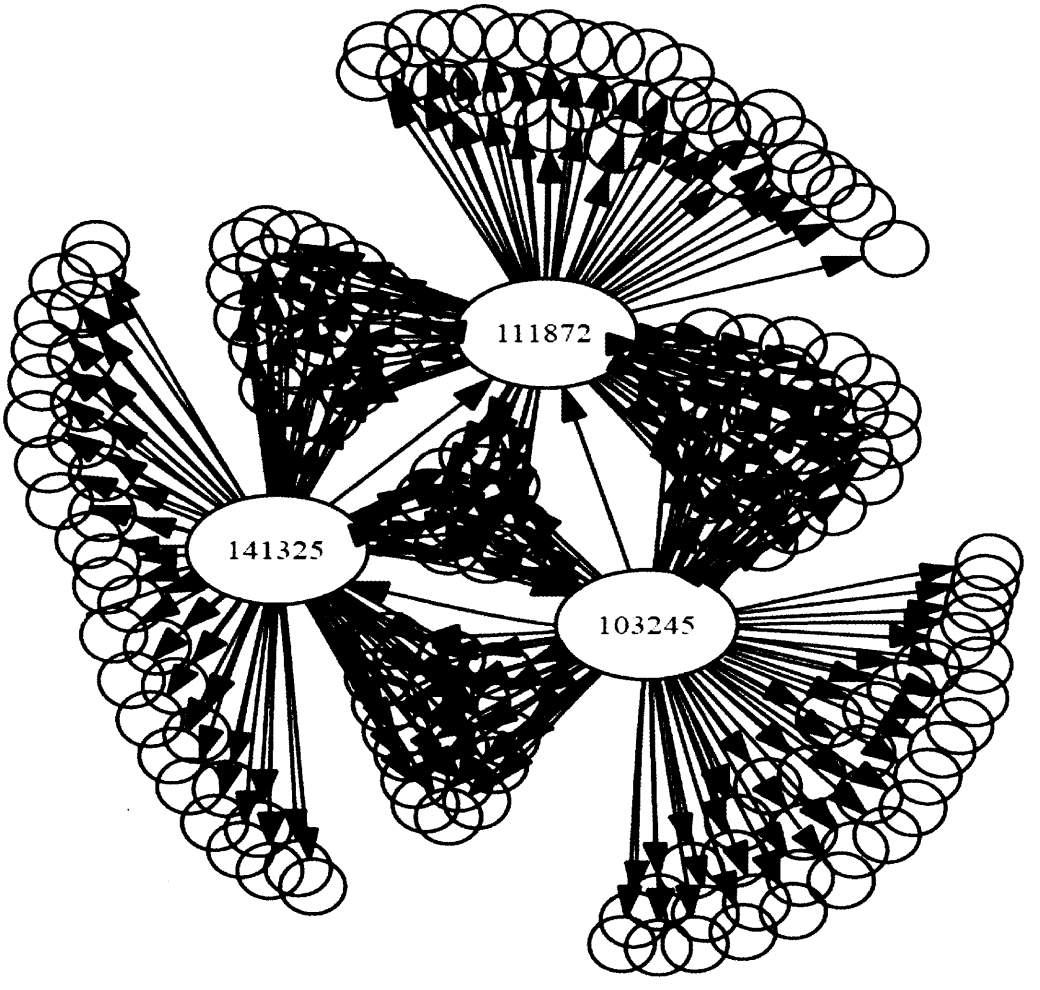


Figure 5.8: The visualization of one spam cluster including 3 sites with shared nodes

ID	URL
111872	Allanahand transsexual.2go-porn.com/
141325	amateur-transvestites.2go-porn.com/
103245	alaska-transexual-escorts.2go-porn.com/

Table 5.9: URLs of targets in presented cluster

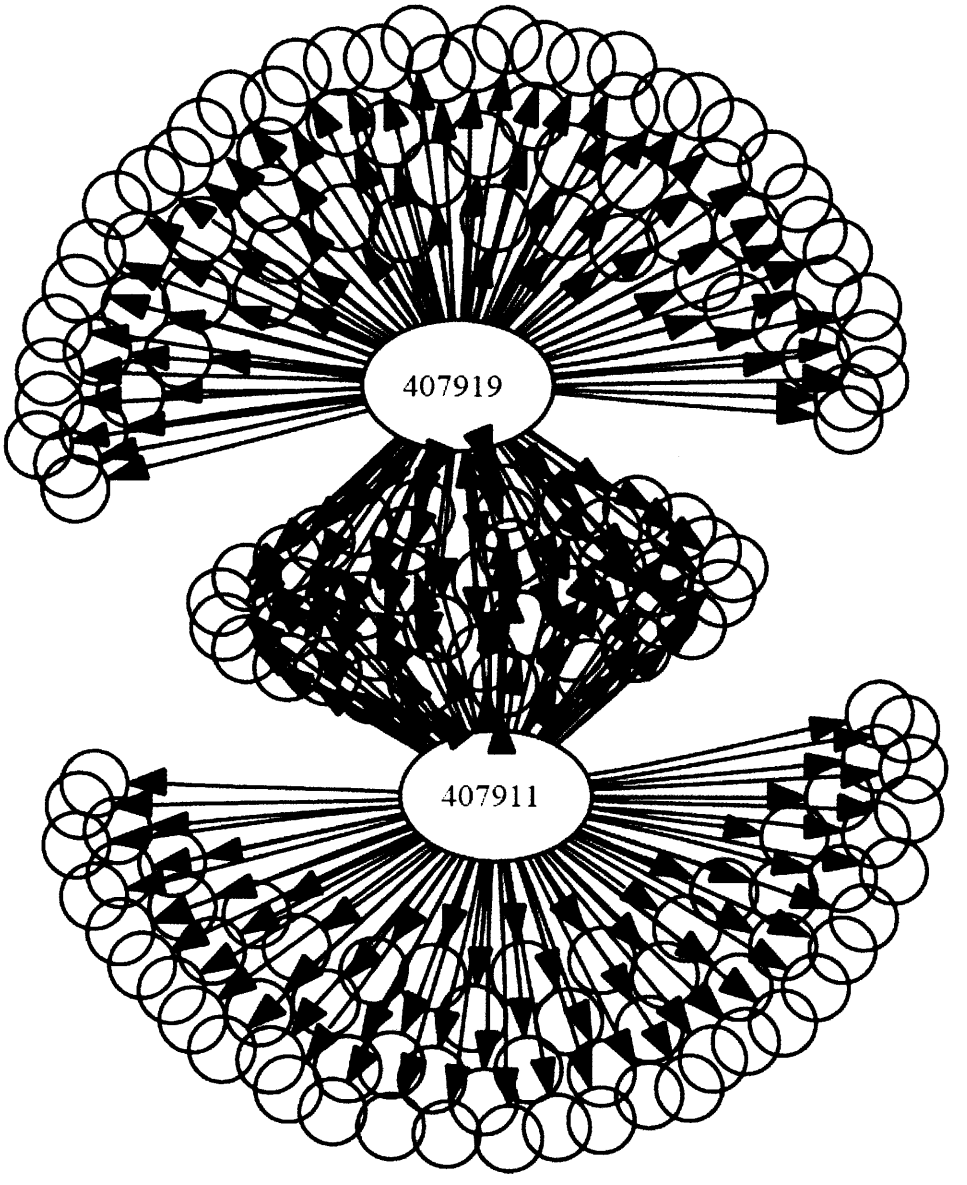


Figure 5.9: The visualization of one spam cluster including 2 sites

ID	URL
407919	button-link.logos-l.de/
407911	button-html.logos-l.de/

Table 5.10: URLs of targets in presented cluster

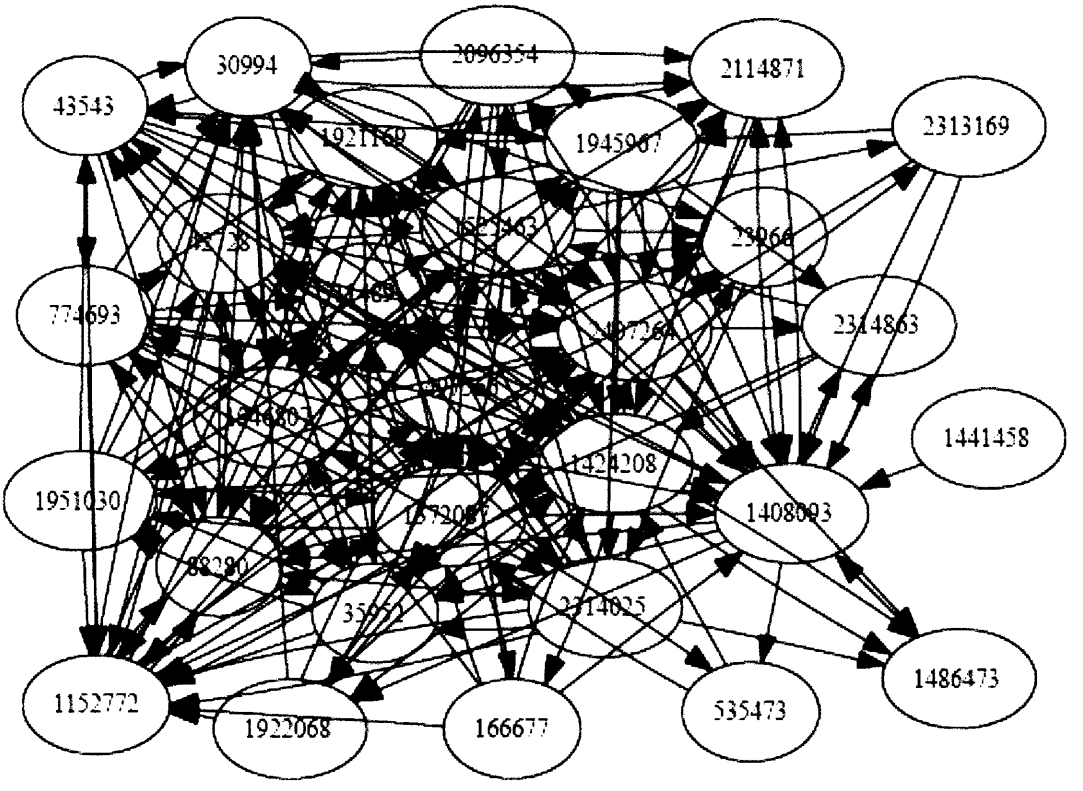


Figure 5.10: The visualization of one spam cluster including 29 sites

ID	URL
43543	timetv.eyejob.co.kr/
774693	eyenude.eyejob.co.kr/
1951030	onkoreasex.eyejob.co.kr/
1152772	hardcore.eyejob.co.kr/
88280	adultv.eyejob.co.kr/
30994	3exdom.eyejob.co.kr/
42728	69sexual.eyejob.co.kr/
1441458	joyitv.eyejob.co.kr/
2096354	popsex.eyejob.co.kr/
1525463	kwaboochon.eyejob.co.kr/

Table 5.11: URLs of part targets in presented cluster

5.8 Analysis of results

We can see from Table 3 and Figure 5 that Pattern Co-citing and Circle provide better performance as the numbers of non-spam sites are comparatively smaller.

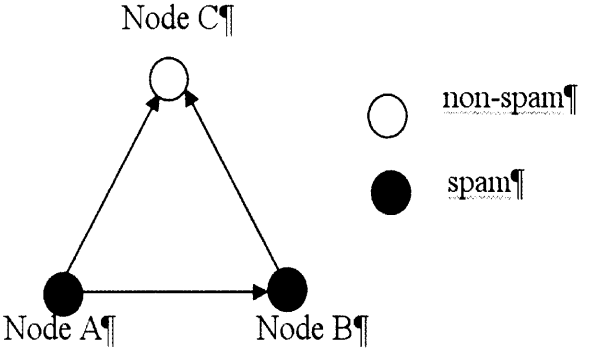


Figure 5.11: A case of spam sites point to non-spam site

A theoretical explanation of the reason is: spam sites could have outlinks pointing to non-spam sites to boost hub values as shown in Figure 6. In this case, node A and B are Co-citing to node C while node B and C are Co-cited by node A. Therefore, non-spam sites which are included in the clusters based on edge (A, B) should be smaller than based on edge (A, C) or (B, C).

Spam sites, which are generally accepted to be automatically generated, have more circle-links than the non-spam sites, which are authored by humans. Therefore, pattern Circle has good performance in detecting spam sites.

Chapter 6

Conclusion

This paper presented a technique to detect web spam from a densely connected directed graph of sites. By applying union-find algorithm and clustering based on 4 basic patterns, we are able to identify link spams efficiently. Our experimental results demonstrated that we can identify most of link spams. Furthermore, pattern co-citing and pattern circle have better performance to avoid mistaking non-spam sites for spam sites.

References

[1] http://www.graphviz.org/ Information:

[2] E. Amitay, D.Cammel, A. Darlow, R.Lempel, and S. Soffer. "The connectivity Sonar: Detecting site functionality patterns". In Proceedings of the 14th ACM Conference on Hypertext and Hypermedia (HT), Nottingham, UK, August, 26-30, 2003

[3] Andras A. Benczur, K. Csalogany, T. Sartos and M. Uher. "SpamRank-Fully Automatic Link Spam Detection". In 1st International Workshop on Adversarial Information Retrieval on the Web, 2005.

[4] R.Baeza-Yates, C. Castillo, and V. Lopez. "PageRank increase under different collusion topologies". In Proceedings of International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005

[5] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, R. Baeza-Yates. "Using Rank Propagation and Probabilistic Counting for Link-Based Spam Detection". The future of Web Search *Workshop*, May 19-20, Barcelona, Spain, 2006

[6] M. Bianchini, M. Gori and F. Scarselli "Inside PageRank" ACM Transactions on Internet Technology (TOIT) Volume 5 , sites 92-128, USA ,Feb 2005,

[7] K. Bharat and M. R. henzinger. "Improved algorithms for topic distillation in a hyperlinked environment". In proceedings of SIGIR-98, 21st ACM International Conference on Resare h and Development in Information Retrieval, sites 104-111, Melbourne, AU, 1998

[8] K. Bharat, B. Chang, M. Henzinger and Ruhl, "Who links to whom: Mining linkage between web sites". In Proceeding of the IEEE International Conference on Data Mining (ICDM) 2001

[9]S. Chakarbati, B.E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A, Tomkins, D, Gibson and J. M. Kleinberg. "Mining the Web' link structure. Computer", 32(8): 60-67, 1999

[10] B.D. Davison. "Recognizing nepotistic links on the web". In AAAI-2000 Workshop on Artificial Intelligence for Web search, sites 23-28, Austin, USA, July 30, 2000

[11] C. Ding, X, He, P. Husbands, H. Zha and H. Simon. "PageRank, HITS and a unified framework for link analysis". In Proceedings of the 25th ACM International Conference on Research and Development in Information Retrieval (SIGIR), pp 353-354, Tampere, Finland, 2002

[12] N. Eiron, K. S. McCurley and J. A. Tomlin. "Ranking the web frontier". In Proceedings of the 13th International World Wide Web Conference (WWW). Sites 309-381, New York, NY, USA, 2004.

[13] D. Fetterly, M. Manasse and M. Najork and J. Wiener. "A large-scale study of the evolution of web sites". In Proceedings of the 12th International World Wide Web conference (WWW), Budapest, Hungary, 2003

[14] D. Fetterly, M. Manasse and M. Najork. "Spam, damn spam and statistics-Using statistics to locate spam web sites". In Proceedings of the 7th International Workshop on the Web and database (WEBDB), Paris, France, 2004

[15] Z. Gyongyi and H. Garcia-Monlina. "Web Spam taxonomy". In Proceedings of the 1st Internatonal Workshop on Information Retrieval on the Web (AIRWEB), 2005

[16] Z.Gyongyi and H. Garcia-Molina. "Link Spam Alliances". In Proceedings of International Conference on Very large Database (VLDB), Trondheim, Norway, 2005.

[17] Z. Gyongyi, H. Garcia-Monlina and J. Pedersen. "Combating web spam with TrustRank". In Proceedings of International Conference on Very large Database (VLDB), Toronto, Canada, 2004.

[18] R. Lempel and S. Moran. "The stochastic approach for link-structure analysis (SALSA) and the TKC effect". Computer Networks 33(1-6): 387-401, 2000

[19] L. Getoor and C. P. Diehl. "Link Mining: A Survey". ACM special Interest Group on Knowledge Discovery and Data Mining. Volume 7, Issue 2, December 2005.

[20] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. "Trawling the web for cyber communities". In Proc.8th www conference, 1999.

[21] P. T. Metaxas and J. DeStefano. "Web Spam, Propaganda and Trust". In Proceedings of International Workshop on Adversarial Information Retrieval on the Web (AIRWeb) Chiaba, Japan, 2005

[22] H. Ono, M. Toyoda and M. Kitsuregawa "Identifying Web Spam by Densely Connected Sites and its Statistics in a Japanese Web Snapshot" DESW 2005, Japan.

[23] B. Wu and B. D. Davison. "Identifying Link Farm Spam Sites". WWW2005, Chiba, Japan, May 10-14, 2005

[24] B. Wu, V. Goel and B. D. Davison. "Topical TrustRank: Using Topicality to Combat Web Spam". WWW2006, May 23-26, Edinburgh, Scotland, 2006

Acknowledgements

First of all I would like to express my sincerest appreciation to my advisor Professor Masaru Kitsuregawa, for giving me a chance to do this research. His keenness of insight and judgment, constant confidence and persistence has inspired me a lot throughout my study and will encourage me in my future as well.

Second, I would like to express my grateful thank to Associate Professor Masashi Toyoda and Research Associate Miyuki Nakano for constructive criticism, and incredible patience to guide me through my research.

Third, grateful thanks to all members in Kisturegawa Lab who gave me their support in different ways. I am proud for being a part of this group and really had a very great time together with all of you.

Last but not least, my especial thanks go to all my colleagues and friends. You guys' existences make me feel home although I was far from my real family. Thank you so much for the sharing the friendship and experiences in growing mature together.