Clique-Attacks Detection in Web Search Engine for Spamdexing using K-Clique Percolation Technique

S. K. Jayanthi and S. Sasikala, *Member, IACSIT*

***Abstract*—Search engines make the information retrieval task easier for the users. Highly ranking position in the search engine query results brings great benefits for websites. Some website owners interpret the link architecture to improve ranks. To handle the search engine spam problems, especially link farm spam, clique identification in the network structure would help a lot. This paper proposes a novel strategy to detect the spam based on K-Clique Percolation method. Data collected from website and classified with NaiveBayes Classification algorithm. The suspicious spam sites are analyzed for clique-attacks. Observations and findings were given regarding the spam. Performance of the system seems to be good in terms of accuracy.**

***Index Terms*—Clique, link spam, search engine, ranking, search engine optimization.**

I. Introduction

In the last few years, reaching relevant information through the use of search engines has become pervasive. Several sites on the web has growing portion of their traffic coming from search engines referrals. The increasing importance of search engines to commercial web sites has given rise to a phenomenon web spam, web pages that exist only to mislead search engines. The purpose of a ranking system is to assign a relative rank to a Web page according to the quality of the Web page. Although the quality of a Web page can mean many different things to different users, these qualities tend to make users prefer certain Web pages over others and express their endorsements in the form of hyperlinks. Therefore, it is reasonable to expect that when the vast link structure of the Web is analyzed properly, spam could be detected.

II. Clique and WEB Spamdexing

A clique is a subgraph in which all nodes are mutually connected to each other by edges. Cliques of size 3, 4 and 5 are shown in Fig 1.

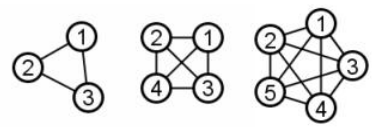

Fig. 1. Clique of size 3, 4 and 5

Manuscript received July 19, 2012; revised August 20, 2012.

S. K. Jayanthi was with Department of Computer Science, Vellalar College for women, Erode-12, India. (e-mail: jayanthiskp@gmail.com).

S. Sasikala was with the Department of Computer Science, KSR College of Arts and Science,Tiruchengode-637211, India (e-mail:sasi_sss123 @rediff.com)

Clique cluster groups the set of nodes that are completely connected to each other. Specifically if connections are added between objects in the order of their distance from one another a cluster if formed when the objects forms a clique. If a web site is considered as a clique, then incoming and outgoing links analysis reveals the cliques existence in web. It means strong interconnection between few websites with mutual link interchange. It improves all websites rank, which participates in the clique cluster. In Fig. 2 one particular case of link spam, link farm spam is portrayed. That figure points one particular node (website) is pointed by so many nodes (websites), this structure gives higher rank for that website as per the PageRank algorithm. Because that algorithm is focussing on the link details.

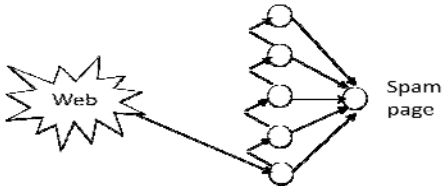

Fig. 2. Link Farm Spam

III. Related Literature

TABLE I: Web spam Literature Review

No	Author name	Title	Focused on	Proposed
1	Baoning Wu and Brian D. Davison	Identifying Link Farm Spam Pages	Link Farms, TKC Effect	Two algorithms. 1. Seed set selection : Common link sets within the incoming and outgoing links of Web pages are considered for selecting the seed set 2. ParentPenalty: Expanding the seed set to include more pages within certain link farms. This spamming page set can be used together with ranking algorithms such as HITS or PageRank to generate new ranking lists.
2	Baoning Wu, Brian D. Davison	Undue Inuence: Eliminating the Impact of Link Plagiarism on Web Search Rankings_	link farm spam, replicated pages and complete hyper-links	1. Algorithm for Finding bipartite components. Based on this concept, they extract one or more complete links from each HTML document and build a document-hyperlink matrix for these pages. The bipartite subgraph within this matrix is a good indication of duplicate pages or link farms. Final adjustment in hyperlink matrix They penalize such links.

				After finding these duplicate complete links they punish them by down-weighting them in the adjacency matrix and then calculate HITS results. Their policy is only to punish links, but not to ban pages containing these links.
3	Zoltan Gyongyi, Hector Garcia Molina	Web Spam Taxonomy	Boosting Techniques, Hiding Techniques, Cloaking and redirection	They presented a variety of commonly used web spamming techniques, and organized them into a taxonomy.
4	Zoltan Gyongyi, Hector Garcia Molina, Jan Pedersen	Combating Web Spam with TrustRank	Spam web pages	TrustRank algorithm for determining the likelihood that pages are reputable. 1. A trust score is calculated 2. Trust propagation 3. Trust Attenuation
5	Jon M. Kleinberg	Authoritative Sources in a Hyperlinked Environment	Search engine	HITS Ranking algorithm Hubs and Authorities
6	Yong Wang, Zhiguang Qin, Bin Tong, and Jing Jin	Link Farm Spam Detection Based on Its Properties	Link Farm Spam	Analyzed the degree distribution and average path length properties of Web network for link farm spam
7.	Qi Chen, Song-Nian Yu, Sisi Cheng	Link Variable TrustRank for Fighting Web Spam	Spam sites	Variance of link structure is taken into consideration, combining with which the ranking scores of websites are judged.
8.	Andras A. Benczur, Dav Siklosi, Jacint Szabo, Iidstvan Biro, Zsolt Fekete, Miklos Kurucz, Attila Pereszlenyi, Simon Racz, Adrienn Szabo	Web Spam: a Survey with Vision for the Archivist	Web archives	Surveyed existing methodology and envision a solution for Web archives to share knowledge and unite efforts in Web spam hunting. They focus on LiWA FP7 project for preventing, detecting and eliminating web spam
9.	Jacob Abernethy, Olivier Chapelle, Carlos Castillo	Graph regularization methods forWeb spam detection	Web graph	WITCH Algorithm, that learns to detect spam hosts or pages on the Web. Unlike most other approaches, it simultaneously exploits the structure of the Web graph as well as page contents and features.
10.	Nikita Spirin, Jiawei Han	Survey on Web Spam Detection: Principles and Algorithms	Principles and Algorithms	They present a systematic review of web spam detection techniques with the focus on algorithms and underlying principles.

IV. Web Search Engine Spamdexing Extraction Using K-CPM Method

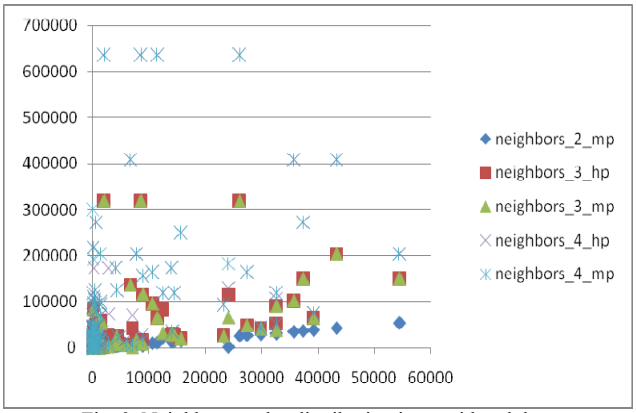

Fig. 3. Neighbour nodes distribution in considered dataset

The details of the neighbour nodes reveal the link farm in many aspects. The distributions of the neighbour nodes in considered dataset are shown in Fig. 3. Maximal cliques are unique and not contained in others. By extracting maximum cliques, the hub of a link-farm can be captured. The steps are given briefly below. And the backtracking has been implied after that to further refine the results.

Step 1: Find the maximal cliques present in website
Step 2: Cluster sites according to the common nodes
Step 3: Sort and Load the edges of the undirected graph
Step 4: If there is a bipartite scenario count the commonly linked nodes. This is complete link exchanges. Two websites mutually share the links and both get higher rank. Consider two websites A and B. Site A contributes all outgoing links to B and B in turn contributes all outgoing links to A.
Step 5: If the number of commonly linked nodes is above a certain threshold THR, then nodes are clustered in the same set.

Finding k-cliques using KCPM

```
CliqueBT (Vector A, j)
{
// If j is equal to size of clique, k, then A is k-clique in the
webgraph
if (j = = sizeClique) then
do
{
numClique++
return
}
else
{
j= j + 1 and if (j <= sizeClique) then do
{
// Let Sj is the set of all candidate vectors
for  j-clique
Sj = getCandidates (A)
} if (Sj is NOT empty)
{
// For each candidate vector in Sj, recursively do
backtracking for k-clique
for (each candidate vector aj in Sj)
do
CliqueBT (aj , j)
}}}
```

V. Dataset

The dataset used in this experiment is collected with information retrieval system which retrieves the inlinks, outlinks and various features listed in Section VI. As a result the numerical values were retrieved from the websites. Samples of five websites were given in Table II. The dataset will be in Excel format which can then be transferred to ARFF format.

After the data collection and preprocessing is over, the classification task is carried out to narrow down the suspicious websites. NaïveBayes Classifier is used here. The suspicious websites are again subject to analysis with utility websphinx and further refined and cliques detected from the dataset.

VI. Clique Extraction and Observations

Analyzing the link properties of website such as indegree (number of inliks), outdegree(number of outlinks), in_out average(for home page and other pages), Pagerank reveals the presence of the spam. The inlinks and outlinks has the intersections (same node acting as inlink and outlink) which is an important aspect to determine the link farm. The Pagerank is considered because it is computed based on the link properties of the website. Totally twelve features are considered and finally the presence of cliques is determined based on the link spamcity score. They are listed below.

1. *F1-avgin_of_out_hp - Average in-degree of out-neighbors of home page (hp)*
2. *F2-avgin_of_out_mp - Average in-degree of out-neighbors of page with maximum PageRank (hp)*
3. *F3-avgout_of_in_hp - Average out-degree of in-neighbors of hp*
4. *F4-avgout_of_in_mp - Average out-degree of in-neighbors of mp*
5. *F5-indegree_hp - Indegree of hp*
6. *F6-indegree_mp - Indegree of mp*
7. *F7-outdegree_hp - Out-degree of hp*
8. *F8-outdegree_mp - Out-degree of mp*
9. *F9-pagerank_hp - PageRank of hp (calculated in the doc graph with no self-loops, using a damping factor of 0.85, with 50 iterations)*
10. *F10-pagerank_mp - PageRank of mp*
11. *F11-trustrank_hp - TrustRank of hp*
12. *F12-trustrank_mp - TrustRank of mp*

Cliques are extracted from the inspected suspicious sites and it is also noticed. Initially the mentioned features are retrieved from the website through web retrieval system. And all the values are processed and consolidated.

With the consolidated data, spamcity score is calculated using Eqn. 1.

$$\text{Link\_Spam\_Score}(WP)=\frac{PR(WP)}{\frac{1}{n}\sum_{wp'?} PR(WP')} \quad (1)$$

A sample calculation of link spam score based on the feature values is given in Table I. Based on the spamcity score the website will be further examined for the presence of link farm in the format of cliques. The link spamicity of a web page is between 0 and 1. The higher the link spamicity, the more the page farm is utilized to boost the PageRank of the target page. If the spamcity score is higher than the threshold α spam is confirmed. Here α=0.5. The navigational pattern of the website remains cyclic and the samples are listed in Table II. Here the numerical values for features are given.

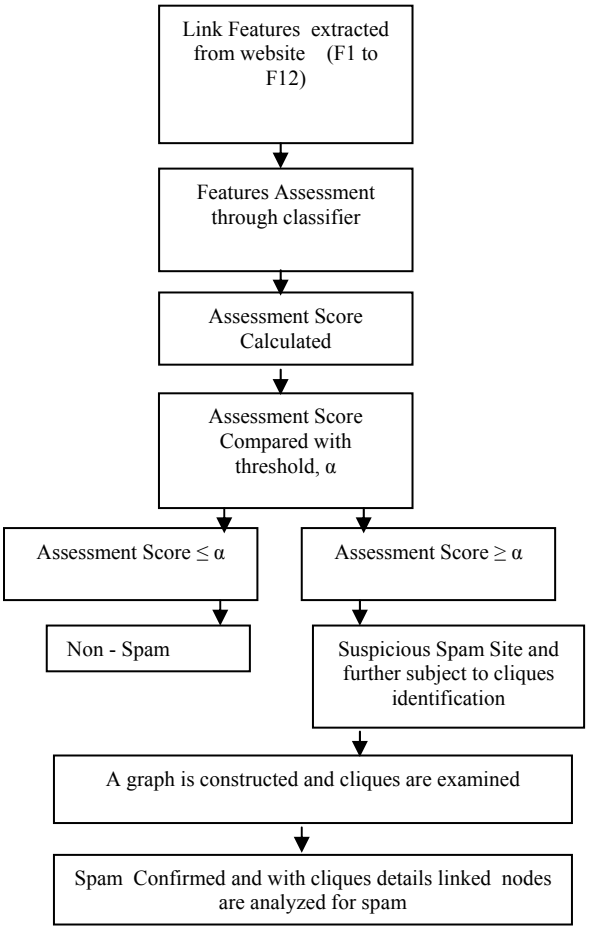

Fig.4. Process of Clique Attacks Detection

Based on the link spam scores the results are further analyzed for the cliques and if it is present then the neighbor nodes are also analyzed for two levels.

Sample values observed while feature extraction are listed in Table II and Fig. 5 represents the in-degree distribution of the considered dataset, Fig. 6 represents the outdegree distribution and Fig. 7 indicates the average of the in and out links in the dataset.

Fig. 4 represents the cliques attacks detection process. Initially the considered dataset is subject to spamcity score calculation and if the score exceeds the threshold then that particular website is analyzed for cliques.

Table II: Features Extracted from Websites

Sno	Host name	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	Link spamcity assessment score
1	109belfast.boys-brigade.org.uk	9.908127	9.908127	5.476464	5.476464	1.386294	1.386294	0	0	19.9363	19.9363	21.2443	21.2443	0.000
2	2accrington.boys-brigade.org.uk	3.178054	3.178054	1.609438	1.609438	0.788457	0.788457	3.781345	3.781345	17.078	17.078	18.4646	18.4646	1.000
3	asami07.cs.rhul.ac.uk	1.94591	1.94591	0.693147	0.693147	2.674149	2.674149	2.824775	2.824775	18.3158	18.3158	19.3123	19.3123	0.500
4	atangledweb.co.uk	4.304065	4.304065	2.70805	2.70805	4.815228	4.815228	2.741711	2.741711	18.2667	18.2667	19.6633	19.6633	0.500
5	beehive.thisisexeter.co.uk	7.986505	7.986505	3.496508	3.496508	10.4057	10.4057	3.273449	3.273449	14.0833	14.0833	14.7952	14.7952	0.750

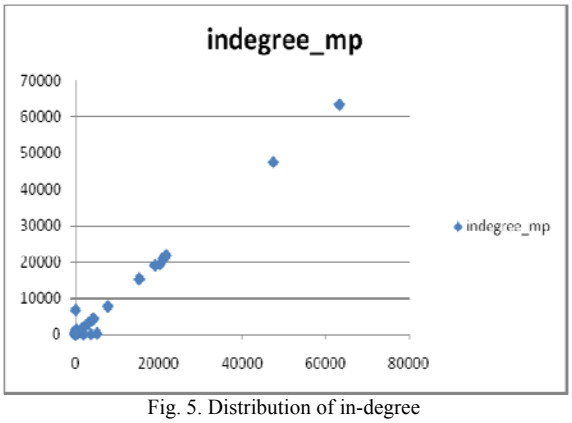

Fig. 5. Distribution of in-degree

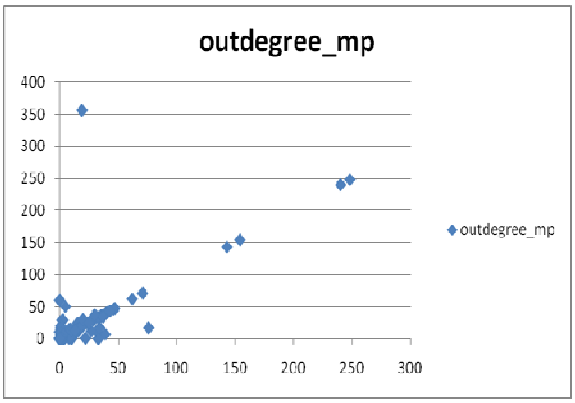

Fig. 6. Distribution of out-degree

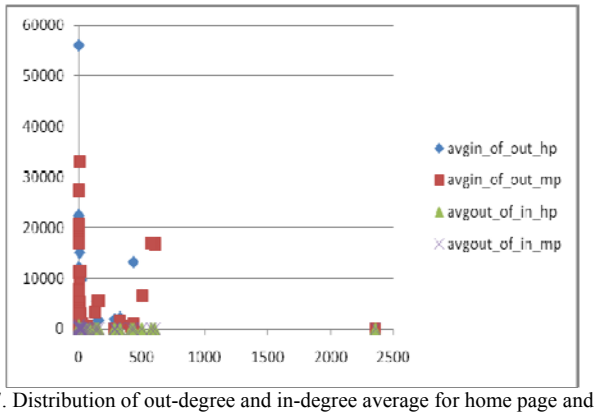

Fig. 7. Distribution of out-degree and in-degree average for home page and other pages

VII. Experimental Results

The results can be outlined as follows. (Due to space constraint only few results have been listed.). Consider classifier C and a test set of documents D. Classifier C computes a label for each document p ∈ D. Let C(p) = 1 if p is classified spam and C(p) = 0 otherwise. Let S ∈ D and N ∈ D denote the set of spam and normal pages (S ∈ N = D). The algorithm categorizes each document in D as spam or a normal page. Let denote SC and NC the set of pages that are detected spam and normal by the algorithm. In summary:

$$S = \{p \in D | \text{ p is spam}\} \quad (2)$$

$$N = \{p \in D | \text{ p is not spam}\} \quad (3)$$

$$SC = \{p \in D | \ C(p) = 1\} \quad (4)$$

$$NC = \{p \in D | \ C(p) = 0\} \quad (5)$$

The following metrics are used for evaluation:

- true positives are the spam documents that are correctly detected as spam, TPC = S∩SC.
- true negatives are the normal documents that are correctly detected as normal, TNC = N ∩ NC.
- false positives are the normal documents that are incorrectly detected as spam, FPC = N ∩ SC.
- false negatives are the spam documents that are incorrectly detected as normal, FNC = S ∩ NC.

The two-by-two matrix of these values forms the confusion matrix:

TABLE III: Evaluation Metrics

	Actual Positive	*Actual Negative*
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

$$\text{Precision} = \frac{tp}{tp + fp} \quad (6)$$

$$\text{Recall} = \frac{tp}{tp + fn} \quad (7)$$

$$\text{True negative rate} = \frac{tn}{tn + fp} \quad (8)$$

$$\text{Accuracy} = \frac{tp - tn}{tp + tn + fp + fn} \quad (9)$$

The two-by-two matrix of these values forms the confusion matrix for the NaïveBayes classifier shown in Table IV.

TABLE IV: Confusion Matrix Generated by NaïveBayes Classifier

=== Confusion Matrix ===		
a	b	<-- classified as
212	10	a = spam
67	3709	b = nonspam

The results generated by the NaiveBayes classifier is given below:

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	3921	98.074 %
Incorrectly Classified Instances	77	1.926 %
Kappa statistic	0.8362	
Mean absolute error	0.0216	
Root mean squared error	0.1334	
Relative absolute error	20.5865 %	
Root relative squared error	58.2658 %	
Coverage of cases (0.95 level)	98.6493 %	
Mean rel. region size (0.95 level)	51.063 %	
Total Number of Instances	3998	

TABLE V: Accuracy Table

=== Detailed Accuracy By Class ===						
TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.955	0.018	0.76	0.955	0.846	0.981	spam
0.982	0.045	0.997	0.982	0.99	0.981	normal
Weighted Avg.						
0.981	0.044	0.984	0.981	0.982	0.981	

As mentioned in the Table IV, the 212 (true positives) websites are taken for the cliques attacks detection and among them many communities strongly interlinking are detected.

Sometimes search engine optimization is applied to genuine websites which may seems like the spam. Those sites are again classified as non-spam in Table VI.

TABLE VI: SPAM CLASSIFICATION IN ANALYZED WEBSITES

Classification Type	Sites	Ratio
Non-Spam	48	24.05
Sales	69	31.96
Pharmaceuticals	3	1.71
Sexually Explicit	92	42.26
Total	212	100%

The assessment score acquired is plotted as a graph in Fig 8. When the assessment score is > 0.5(α) then they are plotted as spam and if it is <0.5(α) it is genuine. For the considered dataset the graph is plotted in Fig 8.

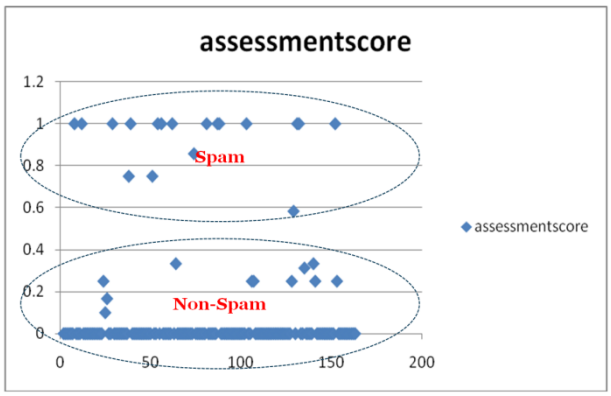

Fig. 8. Assessment score comparison in given dataset

An ROC space is defined by FPR and TPR as x and y axes respectively, which depicts relative trade-offs between true positive (benefits) and false positive (costs). Since TPR is equivalent with sensitivity and FPR is equal to 1 − specificity, the ROC graph is sometimes called the sensitivity vs (1 − specificity) plot. Each prediction result or instance of a confusion matrix represents one point in the ROC space. The Roc curve generated for the naivebayes classifier is given. Here in this experiment AUC for ROC is 0.9809. This is a better result which shows more accurate classification. Fig. 9 shows the ROC curve generated by the NaiveBayes classifier and Fig. 10 shows the cost-benefit analysis for the same.

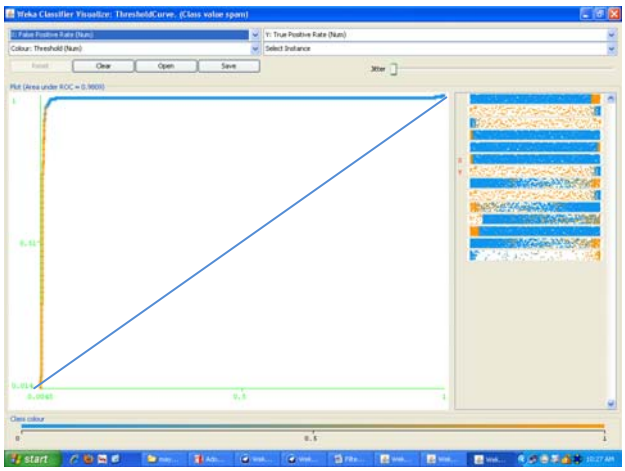

Fig. 9. ROC Curve for NaiveBayes classifier

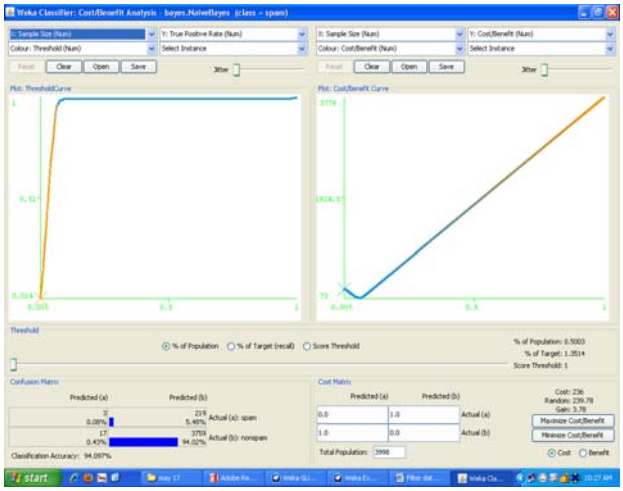

Fig. 10. Cost-benefit analysis for NaiveBayes classifier

Fig. 11 represents a spam website http://www.mobicreed.com where the spam is indicated as red color circle. It actually shows 48,213 links but majority of the links are from the same website. The major link contributor for the website is the same website itself. A cyclic structure is created here.

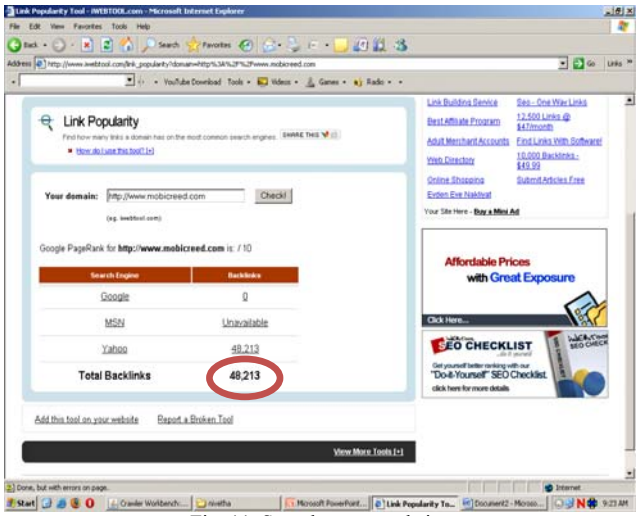

Fig. 11. Sample spam website

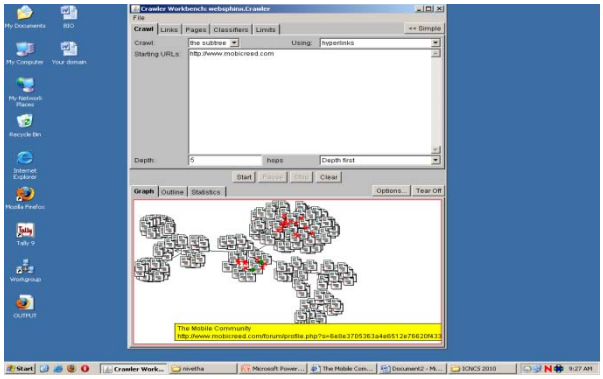

Fig. 12. Spam in hyperlinks- spam is represented as red highlighted

In Fig.12, the websphinx utility is used and it derives the entire graph for the spam website http://www.mobicreed.com. Once a website is classified as spam it will be further subject to the clique attacks. The detection of clique attacks is done through the analyzing the major link contributor website. Here for 212 nodes the analysis is done. This illustrated

example is for clique attack analysis for the mentioned spam site. Fig 13 represents the category classification of the 212 nodes.

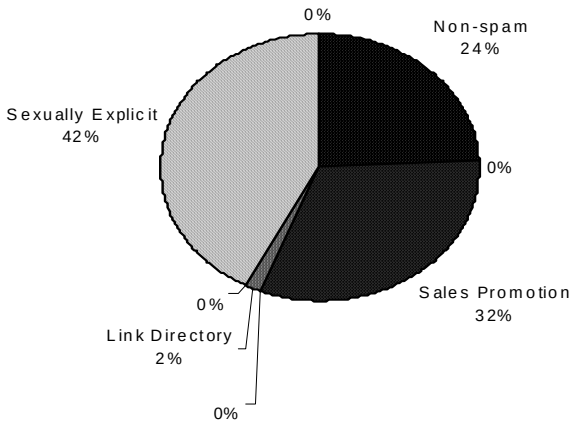

Fig. 13. Ratio of spam classification and maximum cliques in analyzed websites

VIII. Conclusion

The objective of this paper was to classify webspam in given dataset and from the result the cliques are detected. This paper proposes a novel strategy that can be used to filter the spam which possesses the clique attacks. It addresses the link spam type of clique in web search engine. For that the web graph has been created to study the topology of the web site to determine the occurrence of the spam. The inlink, outlink are retrieved from the website with the help of a utility.The suggested approach helps to improve the efficiency in link spam detection by analyzing additional link farms based on constraints as motioned in the algorithm. Based on the recall and precision metrics, conclusion can be made that the proposed system efficiently detects the spam pages. One of the possible improvements is by integrating the weight of web page content relevancy into clique tracking and formulating a collaborative constraint based filter.

References

[1] B. Wu and Brian D. Davison, "Identifying Link Farm Spam Pages," in *Proc.* International *World Wide Web Conference (IW3C2) WWW*, 2005, pp. 820-829.

[2] B. Wu and Brian D. Davison, "Undue Influence: Eliminating the Impact of Link Plagiarism on Web Search Rankings," *Technical Report LU-CSE-06-007, Department of Computer Science and Engineering*, pp. 1-25, 2006.

[3] Z. Gyongyi and Hector Garcia Molina, "Web Spam Taxonomy," *Technical report, Stanford Digital Library Technologies Project*, pp.1-11, 2004.

[4] Z. Gyongyi, Hector Garcia Molina, and Jan Pedersen, "Combating Web Spam with TrustRank," in *Proc. of the 30th VLDB Conference*, 2004, pp.576-587.

[5] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," in *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 1998, pp.1-34.

[6] Y. Wang, Z. Qin, B. Tong, and J. Jin, "Link Farm Spam Detection Based on Its Properties," in *Proc. International Conference on Computational Intelligence and Security*, 2008, pp. 477-480.

[7] Qi Chen, Song-Nian Yu, and S.Cheng, "Link Variable TrustRank for Fighting Web Spam," in *Proc. International Conference on Computer Science and Software Engineering*, 2008, pp. 1004-1007.

[8] A. Benczur, D. Siklosi, J. Szabo, I. Bíro, Z. Fekete, M. Kurucz, A.Pereszlenyi, S. Racz and A. Szabo , "Web Spam: a Survey with Vision for the Archivist," in *Proc. IWAW'08 International Web Archiving Workshop*, 2008.

[9] J.Abernethy, O. Chapelle, C. Castillo, "Graph regularization methods for Web spam detection," *Springer Mach Learn*, pp. 207–225, 2010.

[10] N. Spirin, Jiawei Han, "Survey on Web Spam Detection: Principles and Algorithms," *Newsletter ACM SIGKDD Explorations Newsletter archive,* vol. 13, Issue 2, 2011.

[11] S. K.Jayanthi and S. Sasikala, "Perceiving LinkSpam based on DBSpamClust," in *Proc. IEEE International Conference on Network and Computer Science*, IEEE Press, New York, 2011, pp. 31-35.

[12] S. K. Jayanthi and S. Sasikala, "Link Spam Detection based on Fuzzy C-Means Clustering," *I. J. Next-Generation Networks*, 198, pp. 1-10 2010.

Dr. **S. K. Jayanthi** received the M.Sc., M.Phil, PGDCA, Ph.D in Computer Science from Bharathiar University in 1987, 1988, 1996 and 2007 respectively. She is currently working as an Asso.Professor, Head of the Department of Computer Science in Vellalar College for Women. She secured District First Rank in SSLC under Backward Community. Her research interest includes Image Processing, Pattern Recognition and Fuzzy Systems. She has guided 18 M.Phil Scholars and currently 4 M.Phil Scholars and 4 Ph.D Scholars are pursuing their degree under her supervision. She is a member of ISTE, IEEE and Life Member of Indian Science Congress. She has published 9 papers in International Journals and one paper in National Journal and published an article in Reputed Book. She has presented 21 papers in International level Conferences/Seminars, 20 papers in National level Conferences/Seminars and participated in around 35 Workshops/Seminars/Conferences/FDP.

S. Sasikala, currently working as an Assistant Professor in K.S.R. College of Arts & Science has received the B.Sc(CS) from the Bharathiar University, M.Sc(CS) from the Periyar University, M.C.A. from Periyar University , M. Phil from Periyar University, PGDPM & IR from Alagappa university in 2001, 2003, 2006, 2008 and 2009 respectively. And she is currently pursuing her Ph.D in computer science at Bharathiar University. Her area of Doctoral research is Web mining. She secured University First Rank in M.Sc(CS) Programme under Periyar University and received Gold Medal from Tamilnadu State Governor Dr.RamMohanRao in 2004. She has published 5 papers in International Journals, 10 papers in International Conferences/Seminars, 27 papers are presented in National Conferences/Seminars and acts as a reviewer for 8 journals and 2 conferences and her papers are cited at various publications including IEEE Xplore, International Journals, Wikispaces and Conference Proceedings. She has totally 42 publications and participated in 5 National Conferences/Seminars and 3 workshops and acted as a reviewer for 8 International journals and 2 conferences.