A New Approach to Organize the Results of Searching the Web, using a Combination of Ranking and Genetic Structure-based Clustering

Belal Rostami
Computer Science Department
University of Tabriz
Tabriz, Iran

Shahriar Lotfi
Computer Science Department
University of Tabriz
Tabriz, Iran

ABSTRACT
Web mining means searching the Web for find specific information. Web mining operation should be done in a way to give the best results to the user. Two of the best methods in this area are clustering and ranking Web pages. The hereby-proposed method is a new approach which is a combination of the above-mentioned methods. In the proposed method, first, the Web graph is clustered in two phases, based on structural equivalences; next, each cluster is scored according to its value; then, ranking is done on all present pages in the clusters; and, finally, the final rank of each Web page would be the result of multiplying these two values. In the end, Web pages will be presented to the user based on their final rank. The results obtained from the comparison of the proposed algorithm (GCRM) with other methods indicate a good performance of this algorithm in finding high quality Web pages. Since quality is the main parameter in Web mining, main effort in GCRM algorithm is on increasing the quality of found pages, where, according to the results in this area, GCRM has been successful.

General Terms
Web mining, web graph clustering, web page ranking

Keywords
Web mining, search engines, clustering and ranking

1. INTRODUCTION
Web mining refers to the exploration and search in the Web in order to find specific information and data [22, 26]. Web search engines are responsible for Web search operations but the main problem in Web mining is that for a specific search made by the user, numerous Web pages might be provided, of which usually users only view the first twenty or thirty results presented by the search engines. So, a proper method should be provided in order that the best results are offered to the user. Thus, presenting an efficient method to provide the best results to the user is the main motive in this article.

Generally, in order to search the Web, different Web pages are shown via a graph where nodes represent pages and edges represent links between the pages [4, 5]. In this article a structure-based combinational method is recommended and Web pages are clustered before ranking. Clustering method used in this paper is a new method based on the degree of structural equivalence presented in blockmodeling [2, 3, 7]. In this method there is a 2-phase clustering way where in the first phase, Web graph is clustered with structural equivalence criterion of the links and a representation of every cluster in the second phase enters genetic algorithm to be clustered. In the second phase, with the help of genetic algorithm, the pages are clustered based on the lowest dissimilarity and finally the results of these two phases of clustering are combined with each other. After clustering, a rank is calculated for each present page in each cluster and the pages that are located in desired clusters will get a higher score. Next, the rank of each present Web page in clusters is calculated with PageRank algorithm [17] and in the end final rank of a page would be the result of multiplying the two ranks by each other.

2. THE PROBLEM
Generally a graph is shown by the 2-tuple of G= (V, E). In this formula V is a set of nodes and E is a set of the edges in the form of $E \subseteq V \times V$ [20]. A Web graph is defined in this way that in these graphs, nodes indicate pages and edges represent the links between these pages.

In general, the work done in this research is as follows; first, Web search operation is done and the results which are a number of Web pages along with their contents, relationships and features are entered into the proposed algorithm. So, the input of the problem is a Web graph in which each node contains information about user query and the output of the problem will be corresponding to the user query and based on their quality rate. The purpose in GCRM algorithm is to present the highest-quality results (Web pages) to the user.

3. THE PROBLEM
3.1 Adjacent matrix
The adjacent matrix of G graph with the degree n is shown using a $n \times n$ matrix in the form of $A_G = (a^G_{v,u})$ where the values would be as follows [18]:

$$a^G_{v,u} = \begin{cases} 1, & if\ \{v,u\} \in E \\ 0, & otherwise \end{cases} \quad (1)$$

3.2 Blockmodeling
Blockmodeling is a fundamental problem in network analysis which tries to discover the clusters that have considerable patterns of relation equivalence and the base of blockmodeling process is on similarities [1, 12, and 15].

3.3 Structural equivalence
X and Y pages have structural equivalence when they are connected to the rest of the network in a similar way and fulfill the following conditions [15, 29]:

$$\begin{aligned} &S1: XMY \Leftrightarrow YMX \quad r_{ij} = r_{ji} \\ &S2: XMX \Leftrightarrow YMY \quad r_{ii} = r_{jj} \\ &S3: \exists Z \in U\backslash\{X, Y\} : (XMZ \Leftrightarrow YMZ) \quad \forall k \neq i,j \quad r_{ik} = r_{jk} \\ &S4: \exists Z \in U\backslash\{X, Y\} : (ZMX \Leftrightarrow ZMY) \quad \forall k \neq i,j \quad r_{ki} = r_{kj} \end{aligned} \quad (2)$$

3.4 Dissimilarity degree

If X_i and X_j were two pages among all present pages in a Web graph, then, dissimilarity degree between these two pages would be calculated in the following way [1]:

$$d(X_i, X_j) = \sqrt{(r_{ii} - r_{jj})^2 + (r_{ij} - r_{ji})^2 + \sum_{s=1, s\neq i,j}^{n} ((r_{is} - r_{js})^2 + (r_{si} - r_{sj})^2} \quad (3)$$

4. RELATED WORK

In this section related work in the area of the article are reviewed which are divided into three general sections:

4.1 Related work on clustering

Different proposed methods for clustering the graphs in [19] are divided into four general groups. Partitioning (k-mean, k-medoids [8, 14]), hierarchical (chameleon, HCUBE [16, 19]), density-based (dbscan [21]) and gird-based (sting [9, 27]) methods are these four types.

4.2 Related work in the field of ranking

There are two main methods for ranking Web pages. Content-based (TF-IDF, BM25 [11, 28]) and structure-based (PageRank, HostRank, HITS, WPR [10, 13, 17, 23]) methods are two general categories in this field. However, some methods are proposed recently that are based on user behavior and learning [24, 25].

4.3 Related work on the combination of clustering and ranking

Clustering is a common technique in computer science and one of its important usages is in the field of Web mining and in this regard using clustering methods before ranking Web pages will improve the search results. A method which is based on a combination of clustering and ranking called WSR is proposed in [6].

5. PROPOSED ALGORITHM

GCRM is a combinational method based on clustering and ranking which uses link structure of the pages for this purpose. Unfortunately, not much is done in this area while clustering Web pages before ranking will improve the results. In GCRM algorithm, in the first stage, clustering is done through two phases. For this purpose, at first the Web graph is simplified according to the definitions of structure equivalence and then the pages that are more similar to each other are located in single clusters by using the genetic algorithm. In the second stage, ranking the Web pages is done with PageRank algorithm and the final rank of the Web page is obtained.

5.1 The first stage: The proposed method for clustering the Web graph

The main idea in clustering the Web graph in GCRM is retrieved from the definition of structural equivalence [15]. Accordingly, in the first phase of clustering, the Web graph is clustered using the formula No.2, and since after clustering the graph in this phase, the elements located in the same clusters indicate the same structural patterns, therefore, a representative is selected among each cluster and in this way the Web graph is simplified after being clustered in the first phase. Selected nodes of each cluster form a simple graph which will be clustered again in the second phase of the first stage of GCRM algorithm using genetic algorithm and the formula No.3. In order to better express the way a Web graph is clustered, in figure No.1, a simple Web graph with 15 nodes and the links between nodes is illustrated.

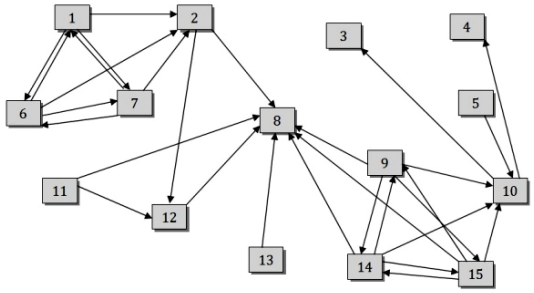

Figure 1: A simple Web graph with 15 nodes

5.1.1 The first phase of Web graph clustering

In the first phase, at first all nodes in the graph are checked in terms of having four conditions presented in the formula No.2, and the nodes that have these four conditions are located in the same clusters. The nodes that are put into a single cluster in the first phase and using this method are completely alike. Figure No.2 shows the Web graph after the first phase clustering.

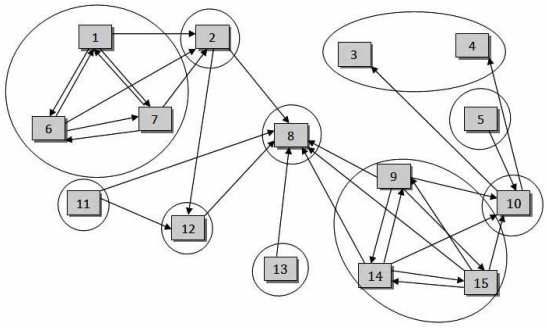

Figure 2: The Web graph after the first phase of clustering with 10 clusters

After clustering the Web graph in the first phase, the elements indicate the same structure, and resulting graph can be simplified. Furthermore, one node from each cluster will enter the second phase of clustering as the representative. In this case, the number of nodes in the simplified graph is equal to the number of clusters in the first phase. After simplifying the Web graph, the resulting nodes will be dissimilar to each other to a certain degree and the amount of their dissimilarity can be calculated with the formula No.3. The dissimilarity matrix of the simplified graph in figure No.2 is shown in figure No.3, where the values indicate that how much a node is dissimilar to the others. This matrix is also used for clustering in the second phase; in this phase and in genetic algorithm the effort is on that the elements which have the least rate of dissimilarity to each other, that is, they are more similar to each other, be in the same clusters.

```
2.2361
2.4495  2.4495
2.4495  2.4495  1.4142
3.4641  3.3166  2.8284  2.8284
3.3166  3.0000  2.6458  2.2361  2.8284
3.3166  3.3166  2.4495  2.4495  2.6458  2.6458
2.6458  1.7321  1.7321  1.7321  2.8284  2.4495  2.8284
2.8284  2.2361  2.0000  2.0000  2.2361  2.6458  3.0000  1.4142
2.4495  2.0000  1.4142  1.4142  2.6458  2.2361  2.6458  1.0000  1.4142
```

Figure 3: Dissimilarity matrix of the simplified graph with 10 nodes

5.1.2 The second phase of Web graph clustering

Genetic algorithm is used in the second phase of clustering GCRM algorithm. The input of genetic algorithm is the dissimilarity matrix of the simplified graph (figure No.3) and the number of its nodes.

- **Coding:** In order to transform the phenotype environment to a genotype one, integer coding is used. In the way that each chromosome of the population is a vector with the length of the number of nodes in simplified graph and the amount of each gene is a number from one to the number of clusters. For instance, figure No.4 shows a chromosome among the population, where the maximum number of the clusters for the simplified graph is considered 3.

1	1	3	2	1	2	1	3	3	3

Figure 4: A typical chromosome of the population for the simplified graph with three clusters

If the number of clusters is considered as c and the number of nodes in a graph is considered n, the size of search space will be c^n, that is:

$$|\text{search space}| = c^n \quad (4)$$

- **Objective function:** Because our purpose in genetic clustering of the simplified Web graph is to put the elements which have the lowest amount of dissimilarity into the same clusters, therefore in the next step, the sum of dissimilarity rates of each node placed in a cluster from each other should be calculated first. If the total numerical value of the dissimilarity in a cluster, like C_s, is considered *Dc*, this value will be calculated with the formula No.5.

$$Dc(C_s) = \sum_{i,j=1,\ i\neq j}^{p} d(i,j) \quad (5)$$

After computing the total dissimilarity of the elements in one cluster, the final sum of these values will be the numerical value of the objective function which is calculated through the formula No.6 and is shown with '*of*'. It should be noted that ch_k in this formula means the chromosome number *k* of the population and n is the number of clusters.

$$of(\mathrm{ch_k}) = \sum_{i=1}^{n} \mathrm{DC(C_i)} \quad (6)$$

The output of genetic algorithm is some clusters of Web pages for a simplified graph which have the lowest rate of dissimilarity to each other. The implementation of genetic algorithm for the first time on simplified graph has created the clustering shown in the table No.1.

Table 1. The result of genetic clustering in the second phase

Cluster No.	Nodes included in the cluster
1	2, 4, 7
2	5, 8, 12, 13
3	10, 11, 15

5.2 The second stage: Simplified graph expansion

After clustering is completed in the second phase, the Web graph should be expanded in order to produce the final clustered graph. In this stage all nodes are examined and the clusters, in which, each node has more than one element in the first phase, are expanded. Final clustering for the mentioned example will be in the form of figure No.5.

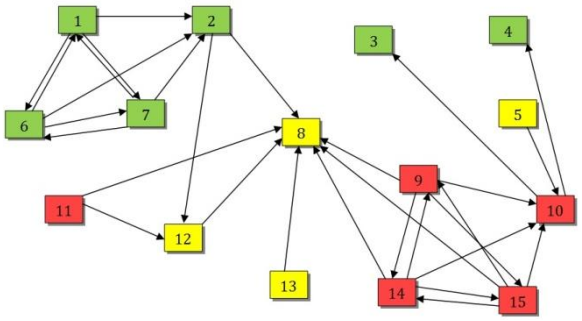

Figure 5: Final clustering obtained after Web graph expansion

5.3 The third stage: Prioritizing the clusters

In this stage, the priority of the clusters is determined based on density concepts. For determining the priority of the clusters, the concept of desired clusters discussed in [18] is used that from this point of view, a desired cluster is the one that has more internal and less external relations. Therefore, the relation rate of each cluster is counted and different clusters are prioritized according to these elements. After calculating the density of each cluster, the clusters get a score on that and this score is applied on each page in a cluster.

5.4 The fourth stage: Ranking the pages

In this stage, the rank of all present pages in the Web graph is calculated with PageRank algorithm.

5.5 The last stage: Final ranking of the pages

In this stage final rank of each page is determined which is the result of multiplying the score of each page on behalf of the cluster by the rank calculated in the previous stage. In the end a page will have a higher rank that is located in the best cluster and has obtained the highest rank in the fourth stage. In this stage, after calculating the final rank, pages are sorted out according to their final rank and are sent to the output.

In GCRM, the quality of a page is determined based on the structure and over two stages of clustering and ranking; this action will eventually improve the results of Web mining. The reason is that Web pages follow similar structural patterns according to their qualities. General framework of GCRM is as follows:

Algorithm: GCRM (Genetic Clustering and Ranking Method)

//Step of GCRM

1. Clustering web graph based on structure

Phase 1: Clustering web graph using structural equivalence and simplify web graph

Phase 2: Clustering the simplified web graph using GA

2. Expand the simple web graph to create final clusters

3. Compute score of clusters

4. Ranking web pages using PageRank

5. Compute the final rank and sort and send to output

6. COMPARISON AND EVALUATION

GCRM is implemented in MATLAB programming environment and in order to compare it with the previous approaches, benchmarks of newsgroups are used that include thousands of Web documents on different fields.

6.1 Evaluation criteria

In this section the main basis of comparison is examining the quality of the found pages with GCRM algorithm with other algorithms. For comparing the quality of found pages the following formula is used [25]:

$$P@n = \frac{\#\ of\ relevent\ docs\ in\ top\ n\ result}{n} \quad (7)$$

The above formula shows the ratio of the number of relevant pages found to the total number of pages. Normally, for comparison, the value of n is considered as 20 to 30 because usually users observe the first 20 to 30 search results. In general, for each query entered by the user, there are three kinds of pages on the Web based on their quality. These pages are called hot, medium and cold which contain high quality, medium quality and low quality information, respectively. An algorithm that takes more hot pages to the output in the first offered n results will have a better performance in Web mining. In order to evaluate the proposed algorithm, two sets of benchmark with "Band" and "Goat" queries are used.

6.2 Comparing GCRM with PageRank in terms of the number of hot pages

In this section, the results of GCRM algorithm are compared with PageRank. The best result and also the average of its responses in 10 times of running are shown in table No.2 for the query of "Band" and in table No.3 for the query of "Goat".

Table 2. The number of hot pages found for the "Band" query

Number of initial pages	PageRank	GCRM (the best)	GCRM (average)
5	0	3	1.6
10	0	6	4.3
15	0	9	6.3
20	1	9	8.3
25	2	9	7.7
30	3	11	9.5

Table 3. The number of hot pages found for the "Goat" query

Number of initial pages	PageRank	GCRM (the best)	GCRM (average)
5	0	5	2.8
10	3	9	6.1
15	5	10	7.6
20	7	13	11.4
25	8	16	14.0
30	9	17	16.0

Considering the numerical values obtained from the comparison of GCRM with PageRank, the efficiency of them is shown in figures 6 and 7 for the web mined query titles.

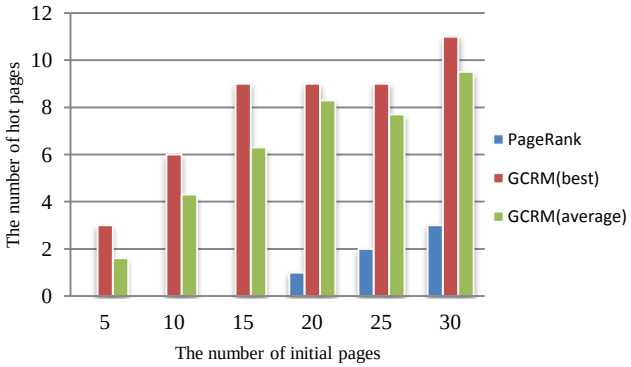

Figure 6: Diagram of the best and average responses obtained through 10 times of running for the "Band" query

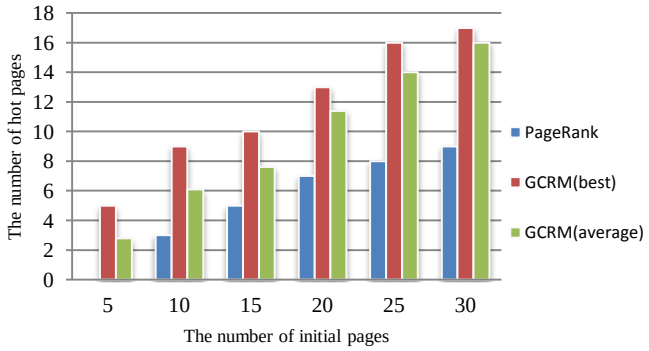

Figure 7: Diagram of the best and average responses obtained through 10 times of running for the "Goat" query

6.3 Comparing GCRM to PageRank in terms of total quality

Because, according to the quality of a Web page, three kinds of pages are presented to the user, comparing general quality of pages is the next comparison subject. In this situation web pages, receive specific coefficients based on their quality (hot, medium, cold) and the amount of general quality of the pages will be equal to the sum of all qualities obtained. Total quality of Web pages obtained from search on the mentioned benchmark in average and the best quality form are shown in the table No.4 with the query of "Band" and in the table No.5 with query of "Goat".

Table 4. Total quality of pages for "Band" query

Number of initial pages	PageRank	GCRM (the best)	GCRM (average)
5	6	11	8.4
10	11	23	19.0
15	16	34	29.6
20	23	39	36.6
25	31	47	43.7
30	39	55	52.7

Table 5. Total quality of pages for "Goat" query

Number of initial pages	PageRank	GCRM (the best)	GCRM (average)
5	5	15	10.8
10	16	26	22.0
15	25	36	30.5
20	34	46	41.5
25	41	57	53.3
30	48	69	62.0

Regarding the numerical values obtained from the comparison of GCRM to PageRank in term of the total quality of the pages, the efficiency of this algorithm compared to PageRank is shown in diagrams of the figures 6 and 7. As can be seen in these diagrams, GCRM shows a better performance in terms of total quality of Web pages, compared to PageRank.

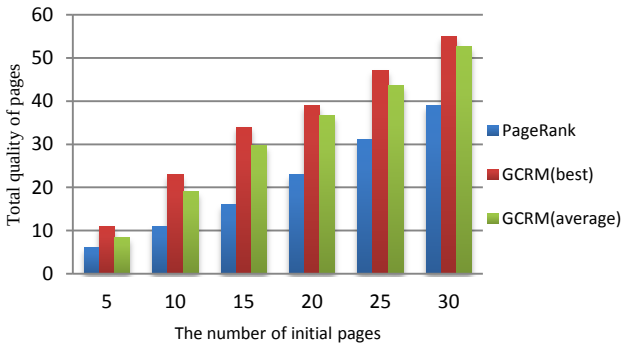

Figure 8: Diagram of the best and average total quality of the responses obtained through 10 times

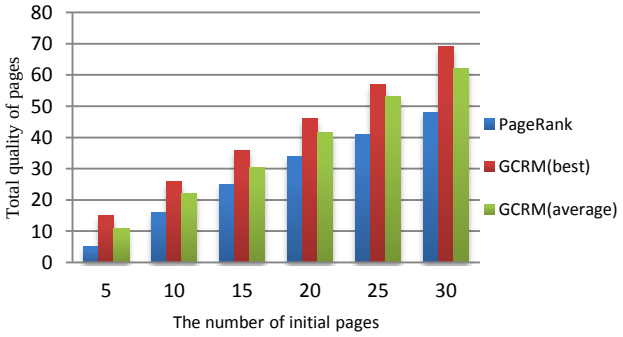

Figure 9: Diagram of the best and average total quality of the responses obtained through 10 times

6.4 Comparing GCRM with other algorithms

In this section the efficiency of GCRM method comparing to other methods is reviewed in terms of P@n criterion. In this comparison the average number of hot pages found in 10 times of running for GCRM is compared to others. The results of examining this comparison are shown in table 6 and figure 10.

Table 6. Comparing GCRM with other algorithms in P@n scale

Number of initial pages	PageRank	WPR	GCRM (average)	DistanceRank
5	0	2	2.8	4
10	3	4	6.1	7
15	5	7	7.6	13
20	7	9	11.4	18
25	8	13	14	21
30	9	13	16	26

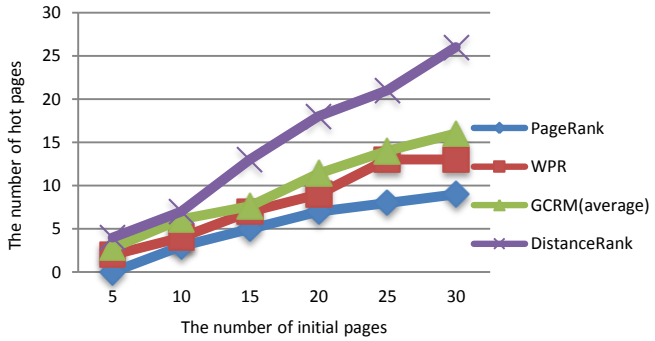

Figure 10: Diagram of comparing GCRM with

6.5 Consistency of GCRM algorithm

In this part, consistency of GCRM is reviewed. In the charts No.11 and 12 the consistency of the proposed algorithm are shown with "Band" and "Goat" query. In this chart, the horizontal axis shows the number of 10 runs and the vertical axis indicates the number of present hot pages in the first 25 found pages. Also for a more precise review, the average and standard deviation of the data are also shown on each chart.

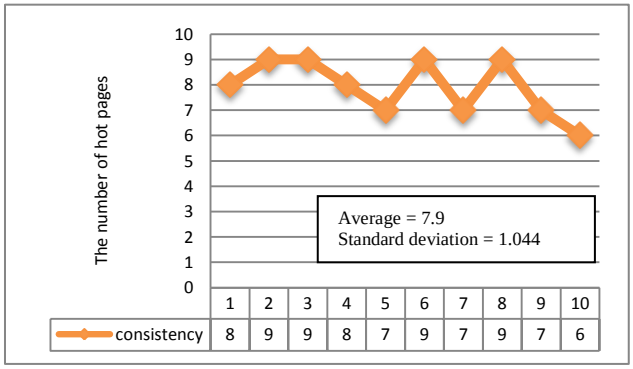

Figure 11: Diagram of GCRM algorithm consistency in 10 runs compared to the number of hot pages found in the first 25 search results on the first benchmark with the "Band" query

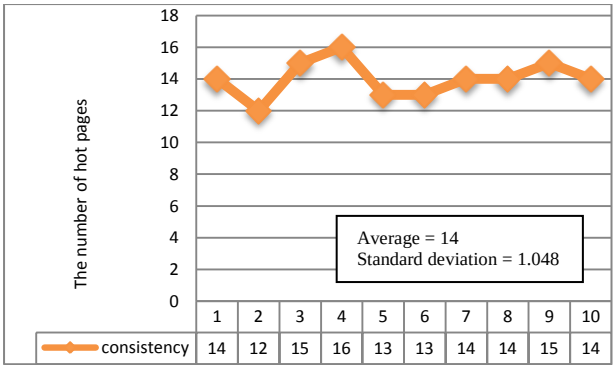

Figure 12: Diagram of GCRM algorithm consistency in 10 runs compared to the number of hot pages found in the first 25 search results on the first benchmark with the "Goat"

6.6 Convergence of GCRM algorithm

In this section, convergence of GCRM algorithm is discussed in the clustering stage (second phase) where genetic algorithm is used. The convergence rate of GCRM is shown in figures 13 and 14. In these charts, horizontal axes express the number of genetic algorithm generations and vertical axes indicate the fitness of the individuals. As can be inferred from these figures, with the repetition of genetic algorithm, the algorithm is converged to better responses.

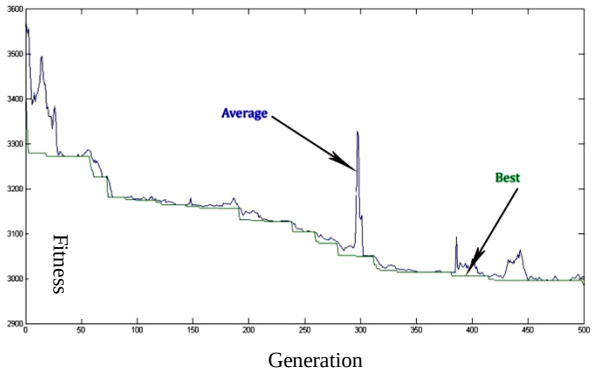

Figure 13: Convergence diagram of GCRM algorithm in run for 500 generations on the first benchmark with "Band" query

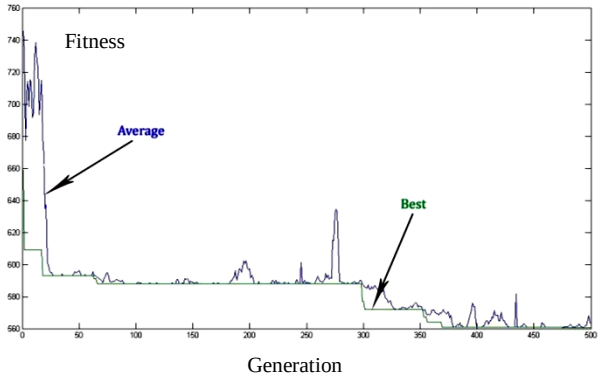

Figure 14: Convergence diagram of GCRM algorithm in run for 500 generations on the first benchmark with "Goat" query

7. CONCLUSION

With the expansion and increasing growth of the Web, the World Wide Web has been transformed into a very important and valuable source of information. Following the rapid growth of the Web, retrieval of the information from this vast source has become a very important challenge in the recent decades. Ranking algorithms are methods offered to retrieve information rapidly and precisely from this world of information. One of the best methods for ranking is structure-based ranking. The main advantage of using structure in ranking is using the content of other pages to determine the rank of a page. A method that will improve ranking on the Web is using clustering methods before applying ranking to the Web pages, which are the main idea and the subject in this paper. The results of comparing GCRM with other methods indicate a good function of this algorithm in finding high quality pages; the main reason of that is using useful link information in clustering stage as well as ranking stage.

8. REFERENCES

[1] Batagelj V., Mrvar A., Ferligoj A. and Doreian P., "Generalized Blockmodeling with Pajek," Metodolo˙ski zvezki , pp. 455-467, 2004.

[2] Batagelj V., "Notes on blockmodeling," Social Network, Vol. 100, No. 19, pp. 143-155, 1997.

[3] Carolyn J. A. and Wasserman S., "Building stochastic blockmodels," Social Networks, pp. 137-161, 1992.

[4] Cason T. P., September 2012, Role Extraction in Networks, PHD Thesis, computer faculty of University catholique de Louvain.

[5] Douglas R. W. and Karl P. R., "Graph and Semigroup Homomorphisms on Networks of Relations," Social Networks, pp. 193-234, 1983.

[6] Duhan N. and Sharma A. K., "A Novel Approach for Organizing Web Search Results using Ranking and Clustering," International Journal of Computer Applications, Vol. 5, No. 10, pp. 8887-8896, 2010.

[7] Faust K. and Wasserman S., "Blockmodels: Interpretation and evaluation," Social Networks, pp. 5-1, 1992.

[8] Guenoche A., "Comparing Recent Methods in Graph Partitioning," Electronic Notes in Discrete Mathematics, Vol. 22, pp. 83–89, 2005.

[9] Grabmeier J. and Rudolph A., "Techniques of Cluster Algorithms in Data Mining," Data Mining and Knowledge Discovery, pp. 303–360, 2002.

[10] Ishii H., Tempo R. and Wei Bai E., "A Web Aggregation Approach for Distributed Randomized PageRank Algorithms," IEEE Transactions on Automatic Control, pp. 1203-1232, 2012.

[11] Jain R. and Purohit G. N., "Page Ranking Algorithms for Web Mining," International Journal of Computer Applications, Vol. 13, No. 5, pp. 8887–8891, 2011

[12] Jessop A., "Blockmodels with Maximum Concentration," European journal of operational research, pp. 56-64, 2008.

[13] Kamvar S., Haveliwala T.and Golub G., "Adaptive Methods for the Computation of PageRank," Linear

Algebra and its Applications, Vol. 386, No. 19, pp. 51–65, 2004.

[14] Kohmot K., Katayama K. and Hiroyuki N., "Performance of a Genetic Algorithm for the Graph Partitioning Problem," Mathematical and Computer Modeling, Vol. 38, pp. 1325–1332, 2003.

[15] Lorrain F. and White H. C., "Structural Equivalence of Individuals in Social Networks," The Journal of Mathematical Sociology, pp. 49-80, 2012.

[16] Murugesan K. and Zhang J., "Hybrid Hierarchical Clustering: an Experimental Analysis," The Journal of Mathematical Sociology, pp. 01-11, 2011.

[17] Page L., "The PageRank Citation Ranking: Bringing Order to the Web," Technical Report, Computer Science Department, Stanford University, 2000.

[18] Schaeffer S. E., "Graph Clustering," Computer Science Review, Vol. 1, pp. 27–64, 2007.

[19] Shaojie Q., Tianrui L., Hong L. and Hongmei C., "A New Blockmodeling based Hierarchical Clustering Algorithm for Web Social Networks," Engineering Applications of Artificial Intelligence, Vol. 10, No. 16, pp. 1-9, 2012.

[20] Tormen C., Leiserson C., Rivest R. and Stein C., *Introduction to Algorithms*, McGraw-Hill, 2001.

[21] Weining Q. and Aoying Z., "Analyzing Popular Clustering Algorithms from Different Viewpoints," Journal of Software, pp. 1382–1392, 2002.

[22] Wu X., Kumar V., Ross Quinlan J. and Ghosh J., "Top 10 Algorithms in Data Mining," Knowl Inf Syst, Vol. 10, No. 1007, pp. 1-37, 2008.

[23] Yan L., Gui G., Du W. and Guo Q., "An Improved PageRank Method based on Genetic Algorithm for Web Search," Procedia Engineering, Vol. 15, No. 34, pp. 2983– 2987, 2011.

[24] Zareh Bidoki A. M. and Yazdani N., "DistanceRank: an Intelligent Ranking Algorithm for Web Pages," Information Processing and Management, Vol. 44, No. 10, pp. 877–892, 2008.

[25] Zareh Bidoki A. M., Oroumchian F., Ghodsnia P. and Yazdani N., "A3CRank: an Adaptive Ranking Method base on Connectivity, Content and Click-through Data," Information Processing & Management, Vol. 46, No. 2, pp. 159-169, 2010.

[26] Zdravko M. and Daniel L., *Data Mining The Web*, Wiley, 2007.

[27] Zhang K., October 2007, Visual Cluster Analysis in Data Mining, PHD Thesis, Department of Computing Division of Information and Communication Sciences of Macquarie University.

[28] Zhang D. and Dong Y., "An Efficient Algorithm to Rank Web Resources," Computer Networks, Vol. 33, No. 6, pp. 449–455, 2000.

[29] Ziberna A., "Evaluation of Direct and Indirect Blockmodeling of Regular Equivalence in Valued Networks by Simulations," Metodološki zvezki, pp. 99-134, 2009.

[30]