# Finding Community Base on Web Graph Clustering

**Alireza Rezaee[1], Fariba Jahandideh Shekalgourabi[2]**

[1] Assistant professor of Department of Mechatronic Engineering, Faculty of New Science And Technologies, University of Tehran, Tehran,IRAN, Email: arrezaee@ut.ac.ir
[2] Student of computer software technology, Hashtgerd University, Hashtgerd, Alborz, IRAN.City: Tehran, Email: Fariba_tegvando@ut.ac.ir

**Abstract**
Search Pointers organize the main part of the application on the Internet. However, because of Information management hardware, high volume of data and word similarities in different fields the most answers to the user s' questions aren`t correct. So the web graph clustering and cluster placement in corresponding answers helps user to achieve his or her intended results. Community (web communities) can be used to generate automated directory services. In this paper the act of clustering has been done by finding the complete bipartite sub- graphs. The sub- graphs form the core of a community or clustering and by extending the core we can attain to the whole clustering .The whole set of graphs in England are 18 million pages and 300 million links.

*Keyword***:** Web, Clustering, Community, Graph, fuzzy.



## 1. Introduction

Due to the exponential growth of information and content available on the web and also a lot of changes in the information, search engines play an important role on the Internet. Currently about 25% of people uses search engines to access the desired site. However, due to the non-structured information and managing information bulky hard (crisp) answers provided by search engines is not necessary and most accurate answers may not be suitable. So now the challenge is to use search engines to find the right answers. Web search through clustering various web searching ambiguous queries grouping the information retrieved by using ambiguous English words and people s' names and accurate information using multi agent beam paths through content-driven Web page clustering technique [1]. Since both the answers and users are unreliable and all results are returned, responses are associated with different degrees of membership, an appropriate application of this category is fuzzy logic. Here we have the word of Mr. Zadeh, the father of fuzzy logic [2] "existing search engines have zero deductive capability. To add a deductive capability to a search engine, the use of fuzzy logic is not an option- it is a necessity". Fuzzy logic can be used in different parts of retrieval systems such as search engines. At first, we examine different main parts of this system. We also introduced a technique called (Retrieval of Web documents using a fuzzy clustering) is suggested that creates the clusters of web documents by using fuzzy hierarchical clustering, it means fuzzy clustering method is offered to construct clusters with uncertain boundaries and allows that one object belongs to overlapping clusters with some membership degree[3]. Information retrieval [4] is divided in two fields, the user and information, and a middle field named retrieval. In Fig. 1 (refer to page 14), In the user field, user`s information needs are expressed that in majority cases it should be expressed according to the language of the retrieval system, or the system makes a model of him and makes his information needs according to the user`s behaviour or the history of his behaviour or a direct information

of his expertise. In the information field the knowledge of documents and data should be modelled and organized. Retrieval field is also the intersection of these two and a match of user`s need and information documents. Information needs are sent to the retrieval field by online questions which are retrieved from whole documents, or May there is a question which is exposed to a flow of information (e.g., News) and makes a filter on related documents and separates them. The phase logic can also be used for modelling and classification of the user's answers and the information will be used for indexing fuzzy. In the case of using the fuzzy logic in search engines various works have been done. Also in [5] a layer of fuzzy decision has been placed in user`s section according to the ontology in Google search engine. In this method the search is based on content instead of word. To enhance the power of search pointers Mr. Choi has used fuzzy way [6].Moreover, to find a fuzzy solution in the system index; the fuzzy parameters have been added. the techniques of expressing user`s query in a The purpose of this paper is to achieve fuzzy clustering web pages into web communities, and consequently in terms of the clustering solutions have been classified However, since the volume was too high and too much time was devoted to it and normal clustering was used and additional Fuzzy logic was for future work.

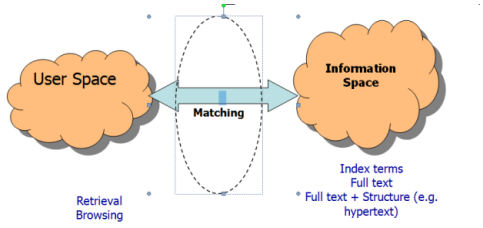


Fig.1. Retrieval system overview

## 2.**Clustering**

The act of clustering is defined as a grouping a set of data without supervision (unlike classification). Documents are usually clustered through their contents which are words. For example, a vector of (n) for each page has been formed and the distance between the vectors of documents are computed and those which have less distance are set in the same group (vector).Word grouping has different applications such as web clustering. For example, we want automatically to groups the entire web pages or a single page into a large site according to their content. If the number of pages is less it`s easily done with this method of clustering (C-means, K-Means). K-means clustering is a commonly used data clustering for unsupervised learning tasks **[7].** But when it is high (millions or billions of pages), the clustering will be very hard and difficult or impossible. Fuzzy C-means clustering (FCM) algorithm, proposed by Bezdek and other researchers can provide an unsupervised approach to the cluster analysis of data [8]. Another application of clustering methods in automatic classification is a search engine query results. Search engines usually ranks a list of URL that contains query terms and are (Boolean search) as its output. For example, the results of the question: "World Cup" into the Google search engine is as follows:

• First and the most important URL is http://www.fifaworldcup.com and the next decade, most of the first 100 addresses are on football (soccer). So "soccer “issue has the largest cluster for this question. • A dozen responses have been ranked http://www.dubaiworldcup.com addresses the issue of horse racing which is different to football. This address can be placed in a separate cluster.

• One of the URL of the response is the soccer robots http://www.robocup.com address that is itself a separate cluster.

• There will be answered, such as sites about Skiing and http://www.skiworldcup.org http://www.fiski.com that needs to be run in a separate cluster. Therefore, the current URL to be placed online, in real-time main clusters based on the cluster restores them. We have some Search engines like http://www.vivisimo.com do this.

The above questions searcher 175 URL with URL titles includes the name and number of spikes in each cluster is returned in brackets as follows:

- Fifa world cup(44)
- Soccer (42)
- Sports (24)
- History (19)
- Rugby world cup(13)
- Women's world cup(10)
- Betting (8)
- Cricket (6)
- Skiing world cup(3)

## 3. Web graph clustering

As previously mentioned clustering web is based on content and high complexity difficult words. Considering the fact that web pages are connected together and form a graph, graph clustering methods would be useful. At present, various methods have been proposed for clustering the web; some of them are briefly explained.

### *2.1. Methods of Bibliographic*

When one page links to another page, certainly the content of two pages are related to each other. But there are a lot of pages that are related to each other (are in a cluster), but not directly linked. Therefore it is needed to find the pages which are

indirectly linked. In Fig.2 two criteria, namely Bibliographic Coupling method by sharing the output link two (multiple) page u and v, and the relationship between them will be determined by the method of Co-citation Coupling of shared pages u and v will be determined by two input links (left) and Co-citation Coupling (right) [9] to obtain indirectly related pages are shown.

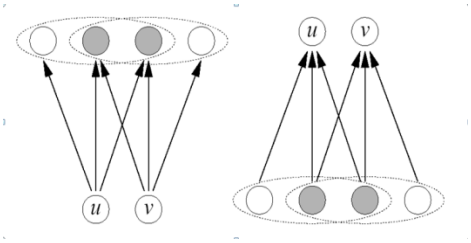


Fig.2. Finding the vertices of Bibliographic coupling (left) and Co-citation coupling (right).

*2.2. Kernels, two-pronged approach (Bipartite Core)*

One of the solutions for a big graph clustering is the use of sub graph, in other words to find pressed sub graphs. In the way of the two-pronged approach [10] the purpose is to find the complete graph. Bisection graph is shown in Klr contains two sets, each of which is represented by l and r, respectively. All members l are connected to all members of r and no connection exists between the members. The split full sub graphs to find the communities in web are useful because of the following reasons: Klr nucleus splits a graph has the property that all vertices in l r with the least amount of Bibliographic Coupling and Co-citation Coupling of all vertices r and l are minimal. In other words, full splits graph has the two properties.

*2.3. Use of Eigen value*

In this way of matrix we compute the Adjacency of the graph, and its Eigen value. Note that n * n matrix with maximum Eigen value n can be used for individual can be calculated in the corresponding cluster [11][12][13]. This method has a high degree of complexity. Also, Mr. William [14] has found a method by using the maximum flows.

## 3. Proposed method

In this paper, kernel method (the complete bipartite graph) is used to find web communities. Test set is Web graph of the (. UK) with 18 million page and 300 million link. [15] Because of the high processing the usual PC is not enough. Test is performed only on the 3 million pages. The goal is to find the core of two parts (i,j) indicates that i represents vertices (fanins) and j represents the right vertices (fanouts) . In other words, the minimum degree of the output vertices and the minimum degree of the input vertices j left is right, i (i = fanin, j = fanout) Clustering algorithm includes the following steps:

*3.1. A simple pruning*

At this stage in a cycle way we travel in the graph and delete all vertices which their output is less than j and input is less than i. In other words, all vertices that could not be in any nuclear will be removed. This operation is repeated until all inappropriate vertices are eliminated.

*3.2. Separating the fan outs*

All vertices which their output at least is j are put in fan-out list.

*3.3. Cutting and nucleus detection*

Choose a vertex x of fan-out list and extract all its output vertices. We also find all inputs of every vertices and take a share of them. If the participant was at least equal to j means that vertex x in a graph is core splits otherwise it`s removed. D) Step c) for all vertices we repeat the fan-out.

## 4. Results

Experiment for different values of the number of fan-out and fan-in community has been done. There are two clustering webs: External and Internal. Creating sites for a large number of pages (up to 3000 pages) might have several clusters. For example, there are a number of research groups that make up a cluster; each cluster is called the inner clusters. In external clustering in a site there are other external clusters. So the inner clustering may consists of the external. To obtain approximate clusters of internal and external processes of the page when the distance between them is less than 1000 links to internal pages And pages of the distance are greater than 3000 is considered as the outer plates. The results in Table 1 for different values of the inner clusters (fanin, fanout) are given. The growth graph charts with full splits (larger fanin and fanout) are less than the number of clusters. Table2 shows the results of clustering external. Expected number of clusters is less than domestically. Table 3 is given. Column of nods the numbers of web pages, addresses column show the address and subject column shows the position. For example, all pages in the first row are concerned to mortgage and insurance. To test the accuracy of several clustering methods and selected address to view the content of the sites.

Table.1
Number of local clusters

| Fan-out | Fan-in | Page range | Number of Internals communities |
|---|---|---|---|
| 2 | 2 | 1000 | 1566 |
| 3 | 2 | 1000 | 667 |
| 3 | 3 | 1000 | 387 |
| 4 | 3 | 1000 | 163 |
| 4 | 4 | 1000 | 127 |
| 5 | 5 | 1000 | 97 |
| 6 | 6 | 1000 | 89 |
| 7 | 7 | 1000 | 24 |
| 8 | 8 | 1000 | 14 |
| 9 | 9 | 1000 | 11 |
| 10 | 10 | 1000 | 6 |
| 11 | 11 | 1000 | 9 |
| 14 | 14 | 1000 | 4 |
| 15 | 15 | 1000 | 3 |

Table.2
Number of foreign clusters based on various levels of fan in and fan-out

| Fan-out | Fan-in | Page range | Number of External communities |
|---|---|---|---|
| 2 | 2 | 3000 | 77 |
| 3 | 3 | 3000 | 18 |
| 4 | 4 | 3000 | 8 |
| 5 | 5 | 3000 | 6 |
| 7 | 7 | 3000 | 2 |

Table.3.
Testing a bunch of pilot

| nodes | Some Addresses | Subject |
|---|---|---|
| 576657<br>986170<br>986171<br>986172<br>986173<br>986174<br>996146<br>1119319<br>1135474<br>1237988<br>1530719 | http://endowment-mortgage.uk-mortgages-advice.co.uk/other.htm<br>http://insurance.lycos.co.uk/PCIS/motorquote/quoterecall.asp<br>http://intu.cem.dur.ac.uk/satis/survey/launch.html<br>http://life-insurance.compare-life-insurance.co.uk/<br>http://mortgage-protection.uk-life-insurance-advice.co.uk/useful-protection-travel.htm | Insurance and mortgage |
| 474429<br>639097<br>986170<br>986171<br>986172<br>986173<br>986174<br>996146<br>1119319<br>1135474<br>1237988<br>1530719 | http://insurance.lycos.co.uk/PCIS/motorquote/quoterecall.asp<br>http://intu.cem.dur.ac.uk/satis/survey/launch.html<br>http://life-insurance.compare-life-insurance.co.uk/<br>http://remedy.telecity.co.uk/telecity/<br>http://debates.by-users.co.uk/<br>http://finance.monster.co.uk/search/ | Insurance and remedy |
| 803474<br>393811<br>414012<br>639013 | http://graduate-jobs-in-wales.ac.uk/english_index.html<br>http://community.monster.co.uk/experts.htm<br>http://content.monster.co.uk/career/networking/party/<br>http://finance.monster.co.uk/ | Jobs , monster |
| 393811<br>414012<br>414031<br>414092<br>414281<br>1079142<br>1584655 | http://community.monster.co.uk/experts.htm<br>http://content.monster.co.uk/career/networking/party/<br>http://leevalley.co.uk/home.html<br>http://sales.centaur.co.uk/remember.asp? | |

## 5. Conclusions and future work

In this report, the web graph clustering methods for finding web communities in the UK (. UK) containing 18 million pages have been done. Due to the high volume of these graphs the use of techniques such as C-Means and K-Means don`t work. So the best way is to find a graph on a compact (dense). Here's the full splits on graphs extracted from the kernel to the rest of the cluster can be obtained. Future Working will be about optimizing the program for clustering large collections, such as the web graph. In addition to adding a Meta search clustering system for advertising (Meta Search Engine) solutions to easily be translated into categories.

## References

[1] Ron Bekkerman, Shlomo Zilberstein, James Allan**,** "Web Page Clustering using Heuristic Search in the Web Graph", Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2007

[2] http://www.cs.berkeley.edu/~nikraves/bisc/sig/internet/ msglaz2.htm. 420-425

[3] Anjali B. Raut, G. R. Bamnote, "Web Document Clustering Using Fuzzy Equivalence Relations", Journal of Emerging Trends in Computing and Information Sciences, CIS Journal, 2010-11.

[4] Ricardo Baeza-Yates and Berthier Ribeiro-Neto., "Modern Information Retrieval", ACM Press/ Addison-Wesley, 1999.

[5] Tomoe Tomiyama, Ryosuke Ohgaya, Akiyoshi Shinmura, Takayuki Kawabata, Tomohiro Takagi, and M. Nikravesh, "Concept-Based Web Communities for Google Search Engine", The IEEE International Conference on Fuzzy Systems, May 25-28, 2003.

[6] Choi, D.-Y., "Enhancing the Power of Web Search Engines by Means of Fuzzy Query", Decision Support Systems, Vol.35, No.1, pp.31-44, 2003.

[7] Chris Ding, Xiaofeng He, "K-means Clustering via Principal Component Analysis", ACM Press, 2004.

[8] V. Loia a,b, W. Pedrycz c,*, S. Senatore, "A P-FCM: a Proximity-Based Fuzzy Clustering for User-Centered web applications, International Journal of Approximate Reasoning, Vol.34, pp.121–144, 2003.

[9] G. W. Flake, S. Lawrence, and C. L. Giles. Ecient Identification of Web Communities", In Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining (ACM SIGKDD-2000), pp.150-160, NewYork: ACM Press, 2000.

[10] S.R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. "Trawling the Web for Emerging Cyber-Communities." In Proceedings of the 8th International World Wide Web Conference, pp.11-16. Amsterdam: Elsevier Science, 1999.

[11] W. E. Donath and A. J. Ho.man. “Lower Bounds for the Partitioning of Graphs.” IBM Journal of Research and Development, Vol.17, 1973.

[12] Jon Kleinberg. “Authoritative Sources in a Hyperlinked Environment”, In Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, pp.668-677, New York: ACM Press, 1998.

[13] S. Brin and L. Page. “Anatomy of a Large-Scale Hypertextual Web Search Engine”, In Proc. 7th International World Wide Web Conference, pp.107-117, New York: ACM Press, 1998.

[14] Gary William Flake, Robert E. Tarjan and Kostas,”Graph Clustering and Minimum Cut Trees”, Internet Mathematics Journal, Vol.1, No.4, Publication Date: 2003/2004.

[15] http://webgraph-data.dsi.unimi.it/