www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2 Issue 4 April, 2013 Page No. 1050 -1053

Graph Clustering In Social Networks Based On Attribute Similarities

S.Kalaichezhian
M.Tech Networking,
Sri Manakula Vinayagar Engineering College, Puducherry
calaichezhian@gmail.com

Abstract
The goal of graph clustering is to partition a large graph into clusters based of node similarities and graph structure. Group of similar data types are done based on some measurements like distance between two vertices, weight of the node, etc. Here service oriented architecture is used in web service to obtain data from different web pages effectively and efficiently. The obtained data will be used to cluster the data based on attribute similarity. The clustered final graph is used to produce a static study about organization. This cluster will help to identify different communication and member's interaction between the community members. This method will help to identify a community in social network or a group of members. In this paper, we are using modified similarity based graph clustering algorithm based on distance or difference between two web pages and weight of the node to group the data set.

1. Introduction

Clustering is a new and developing techniques used for literature studies. Clustering is done to group in similar attributes into one cluster and dissimilar attributes as one cluster. Clustering done in larger graph is mainly done to partition the graph into densely connected components.
This connected component will reveal the several social community and group of members in a social network. Clustering is done in web, social network, telecommunication and biological studies like protein and DNA studies. This graphical

representation is help to identify particular members in a research group or people in a social network, which helps them to share knowledge and ideas between them efficiently and effectively. Same is done in hospital network to identify particular disease infection and spread in an area by obtaining information from hospital network website. The collected details will help to prepare and to identify the area which is more influenced by the deadly disease.

By grouping all organization servers we can obtain efficient data to cluster and to produce a better solution. Here we going to use a well designed web service page by utilizing service oriented architecture method to obtain multiple data from various servers available in internet.

Obtaining Data Using Web Service:

For clustering or community detection efficiently in a social network and in a network, we obtain data sets from internet by getting data from various servers by using service oriented architecture web services. The obtained data is secured and a detail about who can access that information is kept and monitored through this. These data sets are grouped based on similarities and attributes, the clustered graph will help to find information about similar groups in a network and do a static study. This graph also helps group member to share information between them easily.

(I)Parsing obtained information:

The information getting from different servers through internet of various organizations is parsed into single data. The parsed data sets are divided

into groups of similar data sets into one parser. The parsed information is later utilized by modified similarity based graph clustering algorithm to cluster them.

W1	W2	W1	W1	W2	W1	W1

(a) Parsed data sets.

The parsed data sets are further divided into two or more similar data sets obtained from different servers.

(II) Clustering similar data sets:

Graphs are used to denote structural relationship between objects obtained from web and social network applications. Clustering is mainly concentrated in partition of the graph into several connected components. Main goal of graph clustering or community detection is to identifying particular group, disease and different organization members. Most recent method in graph clustering is done based on similarity of attributes.

We proposed an algorithm to model navigational patterns in which, undirected graph is generated. Web pages are used as nodes of the graph and link between the web pages is used as edges of the graph. The degree of connectivity in each pair of pages depends on two main factors: the time position of two pages in a session i.e. time connectivity and the occurrence of two pages in a session is the frequency.

Time Connectivity measures the degree of visit ordering for each two pages in a session that is calculated by the following formula-

$$T_{ab} = \frac{\sum_{i=1}^{n} \frac{T_i}{T_{ab}} * \frac{f_a(k)}{f_b(k)}}{\sum_{i=1}^{n} \frac{T_i}{T_{ab}}}$$

Where Ti is time duration in i-th session, that containing both pages a and b, T_{ab} is the difference between requested time of page a and page b in the session. We consider f(k)=k if web page appears in position k .

Frequency measures the occurrence of two pages in each session that is calculated by the following formula

$$F_{ab} = \frac{N_{ab}}{MAX\{N_a, N_b\}}$$

Where Nab is the number of sessions containing both page a and page b. N_a and N_b are the number of session containing only page a and page b. and this formula also has the values between 0 and 1.

In our approach, Time Connectivity and Frequency plays a major role for finding the degree of connectivity for each pair of web pages and having same importance. We use the harmonic mean of Time Connectivity and Frequency to represent the connectivity of each two web pages, shown as below. We take this formula for weight of each edge in the undirected graph by the following formula

$$W_{ab} = \frac{2 * TC_{ab} * FC_{ab}}{TC_{ab} + FC_{ab}}$$

The data structure can be used to store the weights is an adjacency matrix M where each entry M_{ab} contains the value W_{ab} computed according to above given formula .To limit the number of edge in such graph ,element of M_{ab} whose value is less than a threshold are to little correlated and thus removed. This threshold is named as Minimum Frequency in this contribution.

This proposed clustering algorithm is more efficient in finding the connected components in a graph which has a performance of O(α(n)) amortized time per operation, where α(n) is a very slowly growing function. A cluster-find algorithm that keeps track of a partitioning of a set of nodes/elements over time as certain operations is done. The three main operations are**:(i)Create Set:** Create a new partition containing a single given element.
(ii)Search set: Figure out which partition a given element is in.**(iii)Cluster Formation:** Merge two partitions into It apply Cluster_Find algorithm that finds groups of strongly correlated pages by partitioning the graph according to its connected components. a single partition.

Algorithm

Algorithm: Cluster_Find

Step 1:

for each (a, b)

Calculate

$$W_{ab} = \frac{2 * TC_{ab} * FC_{ab}}{TC_{ab} + FC_{ab}} \quad \text{---------- (1)}$$

Step 2:

If Edge (u, v) <MinFreq then

Delete (Edge (u, v));

Step 3:

for all vertices V of graph G i.e. (V(G)

V=v_1, v_2, v_3…….v_n

Cluster_Find (G) // Cluster Formation

1 for each v€V(G)
2 do Create_Set (v)
3 for eachedge (u,v)€E(G)
4 do if Search_set (u) ≠ Search_set (v)
5 then Merge (u, v)
6. Cluster = Cluster-1

Step 4:

If cluster[i] <MinClusterSize

Delete (Cluster[i]);

i=i+1

return (Cluster);

The Cluster_Find Algorithm takes preprocessed log files as an input. The web pages are represented by the nodes of the graph and link among the web pages are represented by the edges of the graph. Initially all URL assigned to list of web pages then the weight of each edge in an undirected graph is computed by the formula given in equation (1).Now all the edges in the graph having some weight. The edges that are having weight less than the minimum frequency are removed from the graph. Now apply Cluster_Find Algorithm to find Connected Component on V in the graph G.

The procedure Cluster_Find Algorithm initially places each vertex v in its own set. Then, for each edge (u, v), it unites the sets containing u and v. After all the edges are processed, two vertices are in the same connected component if and only if the corresponding objects are in the same set. While searching using a search engine like Google notice that with each result there is a link titled "Similar Pages". If we click this link, Google displays a list of URLs that are related to the item whose "Similar Pages" link we clicked. While I do not know how search engine like Google particularly determines how pages are related but one approach would be the following:

- Let x be the Web page we are interested in finding related pages for.
- Let S1 be the set of Web pages that x links to.
- Let S2 be the set of Web pages that the Web pages in S1 link to.
- Let S3 be the set of Web pages that the Web pages in S2 link to.
- Let S_k be the set of Web pages that the Web pages in S_{k-1} link to.

All of the Web pages in S1, S2--------- S_k are the related pages for x. Rather than compute the related Web pages on demand; we might choose to create the set of related pages for all Web pages once and to store this relation in a database or some other permanent store. Therefore when a user clicks on the "Similar Pages" link for a search term, we simply query the display to get the links related to this page.Search engines have some sort of database with all of the Web pages it knows about. Each of these Web pages has a set of links. We can compute the set of related Web pages using the following proposed algorithm:

1. For each Web page in the database create a set, placing the single Web page in the set. (After this step completes, if it have n Web pages in the database, it'll have n one-element sets.)
2. For a Web page x in the database, find all of those Web pages it directly links to. Call these linked-to pages S. For each element p in S, union the set containing p with x's set.
3. Repeat step 2 for all Web pages in the database.

After step 3 completes, the Web pages in the database will be partitioned out into related groups.

Advantages Of Proposed System

Clustering will help us to obtain a statistical survey or report about the patient's disease in a hospital environment. It also used to identify the areas and hospitals which require more facility to treat the disease base on its severity. In a social network this clustering method helps us to identify an individual group (or) community and provide service. Main advantage of this graph clustering

method is to share knowledge between people and researchers who are grouped under the same category. It is not possible for an outsider to retrieval of data from this clustering group.

Cluster servers offer a considerable improvement in the performance of the network in terms of speed and reliability. Maintenance is very easy with cluster servers. You can switch one server off while others work. This ensures that your network applications are always available. Cluster servers are easy to configure and manage, not taking up much time on many resources. It reduces single points of failure through Exchange Virtual Server (EVS) failover functionality. And has the ability to perform maintenance and upgrades with limited downtime, and to easily scale up your cluster to a maximum of seven active EVSs.

CONCLUSION AND FUTURE WORK

A Clustering algorithm to cluster user navigation patterns is proposed. It is justified the algorithm by the analytical approach as well as experimental evaluation. It used some evaluation methodology that can be used to evaluate the efficiency of the algorithm. For analytical analysis we analyzed the steps of execution of each statement that shows that this algorithm for finding the clusters is more efficient (in terms of time complexity) than the previous existing clustering algorithms and the experimental results presents that our approach can improve the quality of clustering for user navigation pattern in web usage mining systems and also cost effective. For the future, would perfect the algorithm and apply union by rank and path compression heuristic to improve the performance of the Cluster_Find algorithm and some classification methods for classifying user request. This can be used in WUM based online recommendation systems.

References

[1] Yang Zhau , Hong cheng ,Jeffrey Xu Yu ,the Chinese university of Hong Kong "Graph clustering based on structural /attribute similarities",2009.

[2] Benjamin C.M Fung,Member,IEEE,Thomas Trojer, ,"Service-oriented architecture for high-dimensional private data Mashup" ,2012.

[3] Brian, Sugato Bas ,Raymond Mooney,Department of Computer Sciences, University of Texas at Austin, Austin, TX, 78712 "Semi-supervised Graph Clustering: A Kernel Approach".

[4] Guo-Jun Qi[1], Charu C. Aggarwal[2], Thomas Huang[1], [1]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign,[2]IBM T.J. Watson Research Center " Community Detection with Edge Content in Social Media Networks".

[5] Jure Leskovec ,Stanford University,Kevin J. LangYahoo! Research, Michael W. Mahoney ,Stanford University, "Empirical Comparison of Algorithms for Network Community Detection".

[6] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. "Automatic subspace clustering of high dimensional data for data mining applications". In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98), pages 94–105, Seattle, WA, June 1998.

[7] T. M. Apostol. Calculus, Vol. 1: "One-Variable Calculus, with an Introduction to Linear Algebra", 2nd edition. Wiley, 1967.

[8] L.Botton and Y. Bengio. "Convergence properties of the k-means algorithms. In Advances in Neural Information Processing Systems 7 " (NIPS'94), pages 585–592, Denver, CO, Dec. 1994.

[9] D. Cai, Z. Shao, X. He, X. Yan, and J. Han. "Mining hidden community in heterogeneous social networks" In Proc. Workshop on Link Discovery: Issues, Approaches and Applications" (LinkKDD'05), pages 58–65, Chicago, IL, Aug. 2005.

[10] R. Descartes. "The Geometry of Ren´e Descartes". Dover Publications, 1954.