WEB GRAPH CLUSTERING USING HYPERLINK STRUCTURE

San San Tint[1] and May Yi Aung[2]

[1]Department of Research and Development II, University of Computer Studies, Mandalay, Myanmar
[2]Computer Science, MyatLayWyin Street, Yangon, Myanmar

ABSTRACT

Now, information is useful for every environment in which time similarity is more important case. The most of people are strongly interested in Internet. Web pages in the Internet are linked thorough hyperlinks that contain useful information. By using hyperlinks, web graphs are constructed for time similarity web links in which webs have been seen by users at past. These activities are needed to use for tracing who used the websites for something at the time. So this paper provides the history of users who connect the person started the news. We found that the normalized-cut method with the new similarity metric is particularly effective, as demonstrated on a web log file.

KEYWORDS

Eigenvalue Decomposition, Graph Partitioning, Normalized Cut, Similarity Metric, World Wide Web

1. INTRODUCTION

A graph is suitable for World Wide Web (WWW) navigation. Nodes in a graph represent URLs and edges represent links. The entire cyberspace of the WWW can be seen as one graph — a huge and dynamic growing graph. It is, however, impossible to display this huge graph on the computer screen.

Currently researchers interest in using "site mapping" methods to find the way of constructing a structured geometrical graph. This can point out the user through only a very limited region of cyberspace, and the user can not be help for journey through cyberspace.

Online exploratory visualization approach provides major departure from traditional site-mapping methods. This approach does not predict the geometrical structure of a specific Website (a part of cyberspace), but incrementally calculates and maintains the visualization of a small subset of cyberspace. In other words, following the user's orientation, a sequence of Web sub-graphs is automatically displayed with the smooth animation. This approach helps the user to logically survey the entire cyberspace without knowing structure of cyberspace.

In the real world, graphs are constructed by means of the number of nodes and edges. Many graph drawing algorithms have been developed, but most of them have difficulty dealing with large graphs with thousands of nodes. Clustering graphs is one efficient method to draw large graphs even though other techniques exist, such as fisheye view, hyperbolic geometry and distortion-

oriented presentation. A clustered graph can significantly reduce visual complexity by replacing a set of nodes in a cluster with one abstract node. Moreover, a hierarchically clustered graph can find superimposed structures over the original graph through a recursive clustering process.

The link structure of web graph has recently been developed by web graph. The discovery of web communities can be improved by the web graph in which hyperlinks.

2. RELATED WORKS

There is a wide range of work in graph clustering using Normalized Cut. The contributions are focused on improving the algorithm performance, others on proposing different graph modeling and others on the application of this technique for real-world applications.

The web document clustering problem is graph partitioning and measures the partitioning result using the normalized cut criterion [1]. The approach, combining normalized cut and the scaled Fiedler vector and unbiased algorithm can effectively extract different topics contained in the web-graph. Creating clusters with small size by controlling the threshold can be discouraged in Normalized Cut.

The clusters can be obtained high similarity within clusters and dissimilarity between clusters. In their experiment, after choosing suitable threshold, they obtain the clusters, each with distinct topics.

The solution to the normalized cuts problem subjects to a set of linear constraints of the form $UTx = 0$ [2]. Spectral techniques can reduce an eigenvalue problem as well. The set of constraints to $UTx = b$ enforcing constraints is not robust when the constraints are noisy [3]. Solving normalized cuts problem has the computational complexity but is significantly lower than other approaches. Boykov and Jolly find the image segmentation by computing a min-cut/max-flow on a graph in the last decade. The most popular approach to interactive image segmentation in computer vision and graphics has followed that encodes both the user constraints and pairwise pixel similarity [4]. The investigation of lines was modeled the foreground and background regions[5].

4. PROPOSED SYSTEM ARCHITECTURE

In this section, the detailed design of the system is described with the following explanations to understand clearly.

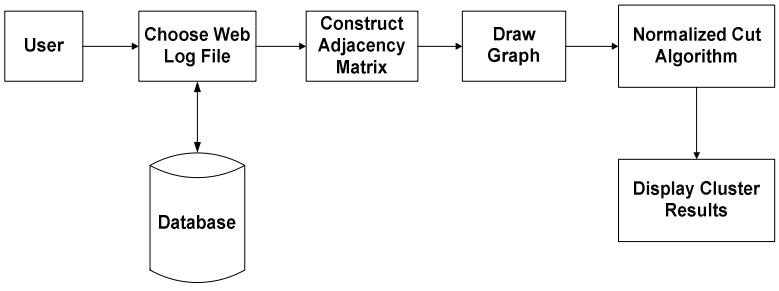

Figure 1. Proposed System Architecture

Figure 1 shows the proposed system architecture. This system is developed for clustering web pages based on the date time information among a set of web pages. In this system, firstly the user first chooses the web log file. The web log file of server is downloaded from a web site. Then these log files are located in the database. After locating, the system gets URLs and Referrers. And then the system computes the adjacency matrix and constructs the graph. After that the system computes the similarity measurement so that the system partitions the web pages by limiting the numbers of clusters. After partitioning, this system uses N-Cut algorithm to compare N-cut value and threshold. And then, this system shows the result by sub-graphs.

4.1. Implementation

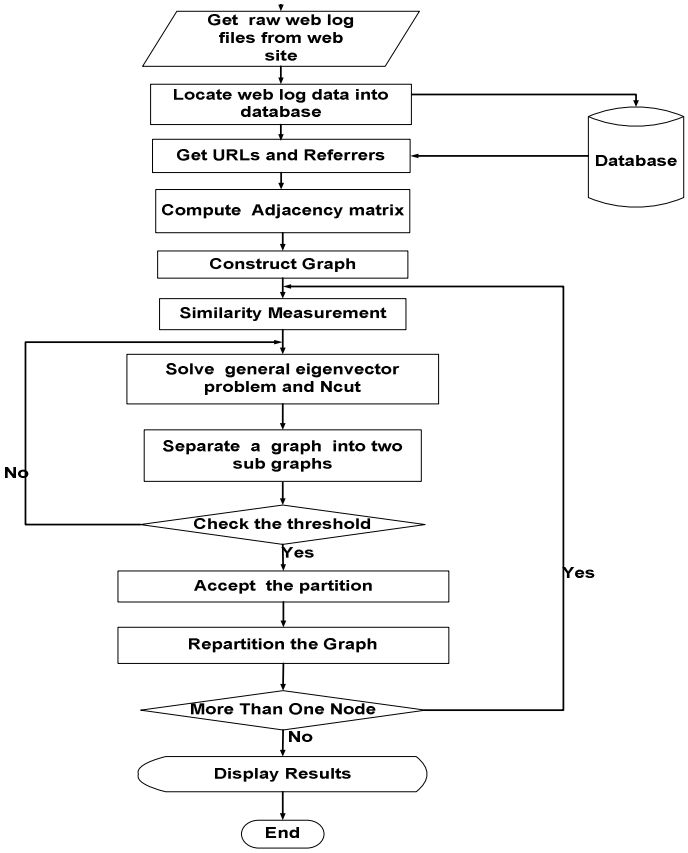

Figure 2. System Flow Diagram

This system focuses on the web graph clustering of web log files according to the different date and time of that file. By applying normalized cut method, web pages can be separated into the clusters of graph .The sample graphs represents time similarity using various web log file data at certain time.

Firstly the raw data of the web log files are downloaded from web sites. After that, these data are located into the database. In the database, data cleaning and transformation are performed. To compute adjacency matrix, the web pages are then retrieved from database. Sample graphs are constructed based on the adjacency matrix and time similarity is measured by using Euclidean Distance method. Ncut is used to solve the eigenvalue and eigenvector of the graph. According to eigenvector corresponding to the second smallest eigenvalues, a graph is separated into sub

graphs. If the value of normalized cut is below a certain threshold, accept the partition and recursively partition the sub graphs, otherwise, stop the clustering. The clusters of the sample graph are displayed as the final result.

4.2. Implementation of the System

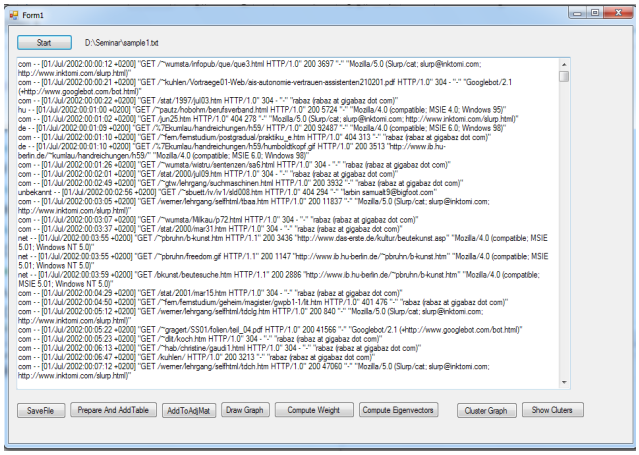

Figure 3. Import Web Log File

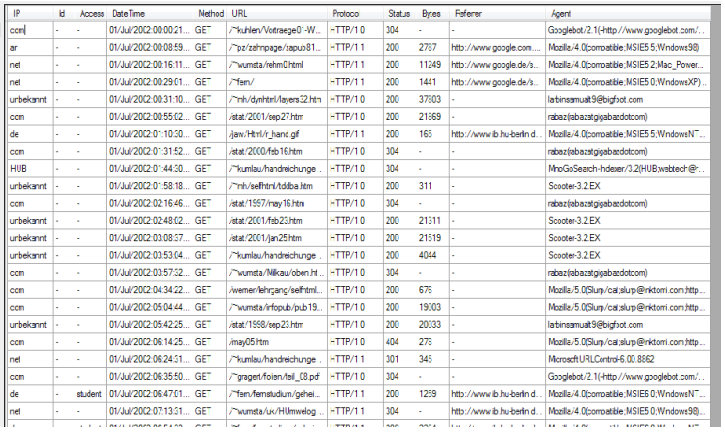

Figure 4. Web Log File DataSet

Figure 3 shows the user import web log file that is extended web log file type. The user can import the raw data. The start button can choose the raw web log file between sample three sample web log files. The user can view this file with a text.

Figure 4 shows the data entry of web log file that consists of IP, Id, Access, Date Time, Method, URL, Protocol, Bytes, Referrer, and Agent. The system used Date Time, URL, and Referrer.

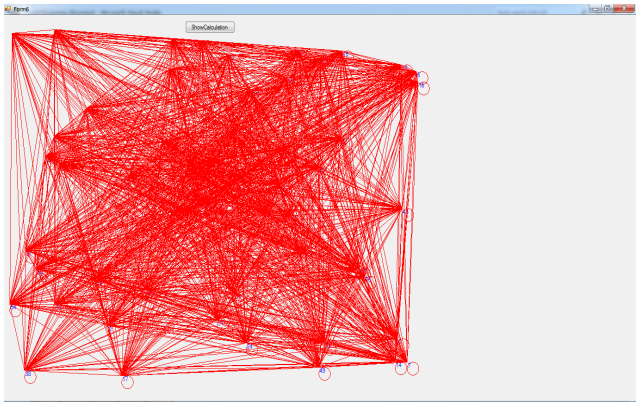

Figure 5. Drawing Graph

Figure 5 shows the graph that consists of nodes and edges. The nodes are web pages and edges are time distance between two web pages.

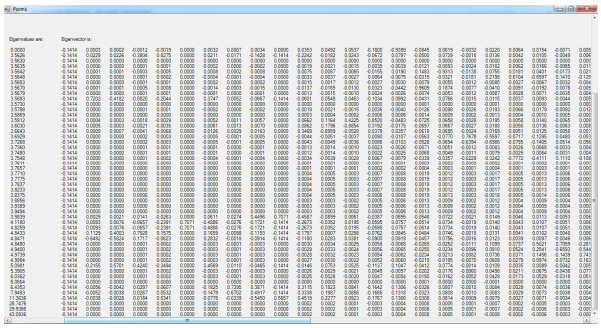

Figure 6. Computations of Eigenvalues and Eigenvectors

Figure 6 shows the computation of eigenvalues and Eigenvector of graph. After constructing weight matrix, diagonal matrix and laplacian matrix are computed. And then, eigenvalue and eigenvector of laplacian matrix are computed to partition the graph. The eigenvalues are n x 1 matrix and the eigenvectors are n x n matrix. The smallest eigenvalue is not used in this system because of zero. The system used the second smallest eigenvector corresponding to the second smallest eigenvalue.

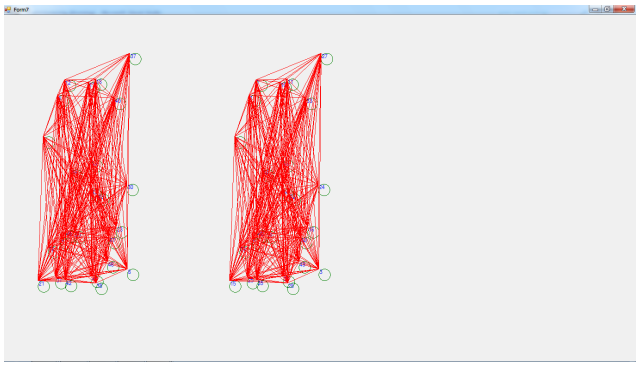

Figure 7. 2 clusters

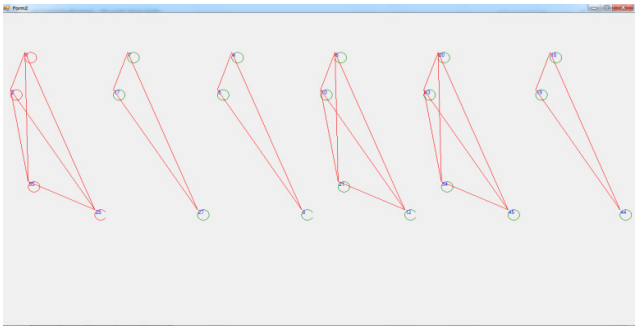

Figure 8. 6 out of 16 clusters

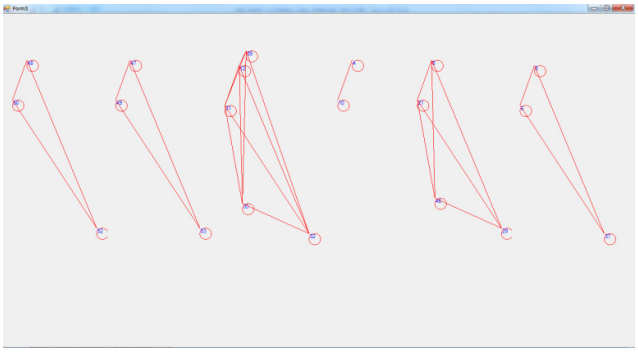

Figure 9. 6 out of 16 clusters

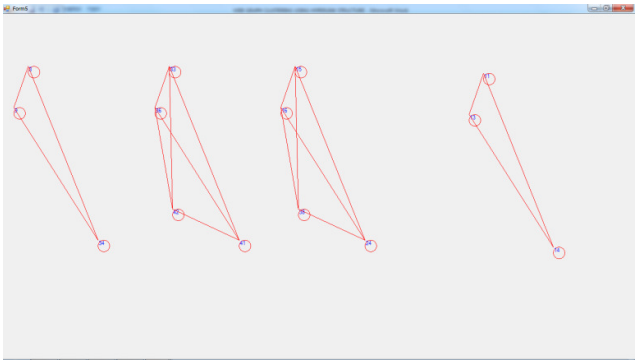

Figure 10. 4 out of 16 clusters

Figure 7, Figure 8, Figure 9 and Figure 10 show the sub graphs of a sample graph. Number of nodes in sample graph is 55 nodes and number of clusters is 16 clusters. The sub graphs show the near time page. The graph is partitioned into sub graphs according to the second smallest eigenvector corresponding to the second smallest eigenvector. The splitting point is zero and the threshold value is the mean value of the eigenvector corresponding to second smallest eigenvalue.

4.5. Experimental Results

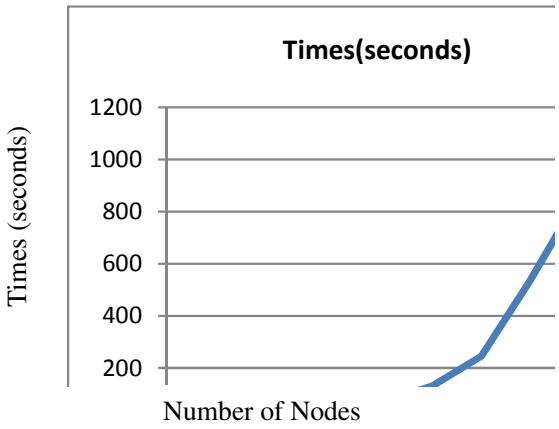

Figure 11. Processing Times

Figure 11 shows the system will take long time as many as nodes from web sites. In figure the number of node is limited to 100 nodes. The more the number of nodes, the more running time will be taken. This system also resolves implementation issues related to clustering technique of Normalized Cut.

5. CONCLUSIONS

This system is implemented as web log analysis that has been widely used to infer near time about web pages. In this system, we present an algorithm to solve the web graph clustering problem. We treat the problem as the graph partitioning, measure the partitioning result using the normalized cut criteria which was first proposed in the field of image segmentation. This algorithm forms a global, unbiased algorithm which can effectively extract different time web page contained in web graph of web log file.

This criterion produces more balanced partitions according to second smallest eigenvalue. The system will produce good results in our experiments and applies into the visualization of a graph. The algorithm is a graph theoretic approach and it thus has wide applications.

This system presents the enhancement of web search using hyperlink structure of the web. In the hyperlink structure, in-links indicate the hyperlinks pointing to a page and the term out-links indicate the hyperlinks found in a page. In this system, Normalized Cut (N-Cut) is used to partition the graph. NCut outputs sample graphs with the second smallest eigenvalue of the Laplacian matrix.

ACKNOWLEDGEMENTS

Our heartfelt thanks go to all people, who support us at the University of Computer Studies, Mandalay, Myanmar. This paper is dedicated to our parents. Our special thanks go to all respectable persons who support for valuable suggestion in this paper.

References

[1] Eriksson, C. Olsson, and F. Kahl, "Normalized cuts revisited: A reformulation for segmentation with linear groupingconstraints", In ICCV, oct. 2007,2058, 2060, 2062.
[2] Y. Y. Boykov and M. P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images", In ICCV, 2001, 2058.
[3] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive image segmentation using an adaptive gmmrf model", In ECCV, pages 428–441, 2004, 2058.
[4] B. Liu, "Web Data Mining", Department of Computer Science, University of Illinois at Chicago, USA.
[5] Abhishek Theory and Social Network Analysis", Awasthi, (2010/2012) "Clustering Algorithms for Anti-Money Laundering Using Graph Autonomous University', Barcelona, MathMods.

Authors

Author is Associate Professor, Head of Department of Research and Development II in University of Computer Studies, Mandalay, Myanmar . Author have worked in University since 1997. Prior to that author spent 6 years as a teacher in Base Education School. Author got B.Sc.(Physics) and M.Sc.(Physics) degrees from Yangon University, Yangon, Myanmar and then M.A.Sc.(Computer Engineering) and Ph.D. (Information Technology) degrees from University of Computer Studies, Yangon, Myanmar. Research on teaching and learning in Cryptography and Network Security, Internetworking with TCP/IP and Digital Fundamental in University.