Citation Network Visualization of CiteSeer Dataset

Afsheen Khalid
Institute of Management Sciences, Peshawar
afdkhan@yahoo.com

Muhammad Tanvir Afzal
Centre for Distributed and Semantic Computing
Muhammad Ali Jinnah University, Islamabad, Pakistan
mafzal@jinnah.edu.pk

Muhammad Abdul Qadir
Centre for Distributed and Semantic Computing
Mohammad Ali Jinnah University, Islamabad, Pakistan
aqadir@jinnah.edu.pk

***Abstract*- Citation networks are very dense and large networks. The visualization of such dense networks is a tough task. Furthermore, most of the available tools provide visualization of such networks in the form of static images. However, the static images are not of any use for the users as user cannot interact with the system and cannot find the required information. In this paper, we have implemented a system that generates a dynamic network of citations. We have used CiteSeer dataset for citation network. We have provided the functionalities where a user can magnify the shown graph and can view all the citations of the required selected paper in a radial tree form. We have also provided a lens facility that provides a hyperbolic view and user can zoom in and out the graph in the area of lens. We created this citation network in Pajek tool and then extended the tool by incorporating JUNG libraries of java.**

I. INTRODUCTION

Scientific research demands the existing research work in interested field in a managed form. Requirements of a researcher vary depending on the role and situation. For example, a new student just entering in the research, might be interested in taking the orientation within the existing work, a reviewer in the correctness and originality checking and an experienced scientist in finding new developments and identifying hot topics in the interested research area. Researchers spend considerable portion of time on these task and want to streamline the process and save time as much as possible. Information visualization is the way through which a researcher can identify the required contextual information instantly from a very huge data set.

Citation networks consist of bibliographical entries. These entries represent scientific works, each being a tuple of attributes such as title, authors, source, date, abstract, keywords, etc. In addition, each entry has a number of references to other entries representing the citations found in the article. Thus, citation networks can be seen as directed graphs where each node represents an article, out edges represent cited papers (i.e. the dependencies of the current paper), and in edges represent citing papers. A citation graph is generally not acyclic since articles may mutually cite each other; this is often the case when an author (or a team of authors) publishes two or more related articles to the same conference. Traditional bibliographical databases generally provide means for searching, sorting, and filtering the citation data in various ways (examples include IEEE Xplore1, the ACM Digital Library [7], and CiteSeer [24]). These database interfaces serve as suitable reference implementations when assessing new visualizations for citation networks. The highly connected and highly contextual nature of citation networks demand that the techniques of information visualization should be used in this area to display the large amounts of data in an interactive way.

Information visualization utilizes computer graphics and interaction to assist humans in solving problems. It is a set of technologies that use visual computing to amplify human cognition with abstract information. Application of information visualization on the computer involves providing means to transform and represent data in a form that allows and encourages human interaction. Data can therefore be analyzed by exploration rather than pure reasoning.

In this paper, we use the citation network of CiteSeer dataset and visualize the cited-by papers information using Pajek and GUESS tools. Due to very dense nature of citation graphs, we use JUNG (Java Universal Network and Graph) library in Java to extract the cited-by papers of a specific paper. We use radial tree layout in which nodes at different levels are kept at circles. Since citation network is very dense and tracing the path for some paper's cited-by papers hides a portion of the graph, we have used the facility of a lens so that any portion of the network can be magnified according to user requirement. Zooming facility is present and network can be zoomed in or out with and without lens.

Below we cover the related work and then in next section the methodology. Results section shows the graph from Pajek and GUESS tools and then how we used to visualize the Pajek graph to show the specific cited-by papers of a target node. The last section summarizes what we have done in this paper and what we plan in future.

II. RELATED WORK

One of the earliest attempts to pictorially represent scientific development was Garfield's istoriograph [1]. This is a diagram of citation patterns depicting the linking of papers forward and backward in time to trace the lineage of ideas over several generations. In a landmark study [2], an historical account of the discovery of the genetic code was correlated with a citation network. Recently these data were reanalyzed using sociometric network analysis [3]. Price's article on citation networks (1965) represented citations between articles as filled cells in a citing/cited matrix. Regularities such as the periodic appearance of review articles, and the distinction between the

archival and research front literatures, were seen as dense vertical or horizontal patterns on the matrix. Price's approach has recently been extended by [4]. The first computer visualization of citation networks was by [5]. This early system allowed forward and backward navigation of citation links, and displayed historiographs on a screen using one-dimensional and multidimensional scaling to fix the positions of documents along the horizontal axis, and publication year along the vertical axis. The most recent development in the graphic display of citation nets is the work of [6] called the butterfly.

JUNG (the Java Universal Network/Graph Framework) [8] is an open source graph modeling and visualization framework written in Java. The framework comes with a number of layout algorithms built in, as well as analysis algorithms such as graph clustering and metrics for node centrality.

JUNG architecture supports a variety of representations of entities and their relations, such as directed and undirected graphs, multi-modal graphs, graphs with parallel edges, and hyper graphs. It provides a mechanism for annotating graphs, entities, and relations with metadata. JUNG also facilitates the creation of analytic tools for complex data sets that can examine the relations between entities as well as the metadata attached to each entity and relation. JUNG includes implementations of a number of algorithms from graph theory, data mining, and social network analysis, such as routines for clustering, decomposition, optimization, random graph generation, statistical analysis, and calculation of network distances, flows, and importance measures.

JUNG provides a visualization framework that makes it easy to construct tools for the interactive exploration of network data. Users can use one of the layout algorithms provided, or use the framework to create their own custom layouts. In addition, filtering mechanisms are provided which allow users to focus their attention, or their algorithms, on specific portions of the graph. It provides powerful APIs for manipulating, analyzing, and visualizing graphs and networks. There exist numerous other packages and tools for visualizing and manipulating networks; e.g. UCINET [9], Pajek [10], R [11] with sna [12], and GFC [13] etc.

Pajek is a program for analysis and visualization of large networks. It has widely used as an efficient analysis tool in all kinds of networks, such as social networks, Internet networks and ISP networks. Pajek provides a wide range of powerful functionalities for network analysis while it is very easy to install and use. Also it is freely available for noncommercial use. All these are the reasons why we choose Pajek as our analysis tool.

GUESS (Graph Exploration System) [14] is an exploratory data analysis and visualization tool for graphs and networks. The system contains a domain-specific embedded language called Gython (an extension of Python, or more specifically Jython) which supports the operators and syntactic sugar necessary for working on graph structures in an intuitive manner. An interactive interpreter binds the text that the user types in the interpreter to the objects being visualized for more useful integration.

Other similar software includes NetMiner[15], StOCNET[16], MultiNet[17], InfoVis [18], InfoVis Cyberinfrastructure [19], Visone [20], Boost [21], JGraph [22], and yFiles [23].

III. METHODOLOGY

We show the citation network of the CiteSeer dataset that has information of more than one million papers. Due to limited system resources, we have visualized cited by information of 13000 papers only but it can be used for the visualization of entire dataset with increased system resources. We used two tables "papers" and "cited_by" of CiteSeer database. Table "cited_by" has random entries about cited by papers-ids. In order to show the cited by papers of a particular paper, we used the SQL queries to extract paper-ids and their cited-by-paper-ids along with year information from "papers" table and "cited_by" table.

Firstly we used the Pajek tool to show the papers and their cited by papers in one graph. Though this graph can be visualized in different layouts, it gives the impression of very dense graph with a number of connecting lines crossing each other even with very small size vertices. Such graphs are not able to give the precise visual information about the connection of cited by papers starting from a particular paper. A user has to look for the cited by papers by tracing the connecting lines. Though these graphs can be zoomed in and out but tracing the cited by papers of some paper that are spanned in the whole graph needs scrolling that can hide the previously scanned papers. In this way, user cannot keep in view the complete chain of papers and the presence of other links and papers make the paths very unclear and noisy.

We made this graph interactive by using JUNG libraries of java in the java platform. It should be noted that JUNG libraries have the support to execute .net files of Pajek tool using Pajek interface. We used this facility to show the complete graph. In order to extract only the cited-by papers of some paper, we used very simple and comprehensive visualization in the form of two panels. In the first panel, the complete graph created by Pajek is visualized and in second panel the sub graph consisting of cited by papers of any paper from the first panel is visualized in radial tree layout format. Circles are used to show the levels of the tree so that all the nodes at one level are placed on one circle. The tree starts with the parent node over which the mouse pointer comes in the first panel and then it is extended with cited by papers over the circles. The sub graph in the second panel keeps changing according to the mouse pointer movement over the nodes in the first panel. Due to this synchronization, we are able to view the cited by papers of any paper in the complete graph in a very neat and clean way. We have also provided a resizable lens in the form of hyperbola that is used to view the graph in 3D plane. Resizing the lens increases or decreases the magnification of graph only in the hyperbolic area. This hyperbolic view can be turned off or on according to user requirement.

Citation networks are cyclic and tree layout does not support cycles. So in the creation of radial tree layout, we got the error on the node that was going to create the cycle. In order to keep all other nodes in the tree form, we just removed these very few

nodes from the chain and informed the user by displaying a message that these papers are also included in the chain at a specific level.

IV. RESULTS

In Pajek, we applied the available layout options on different number of citation records. With less citation records, citation graph is a not so massive but still getting the cited-by papers of a particular paper is not possible e.g. when the Kamada Kawi free algorithm is applied on only 1000 citation records, result is like

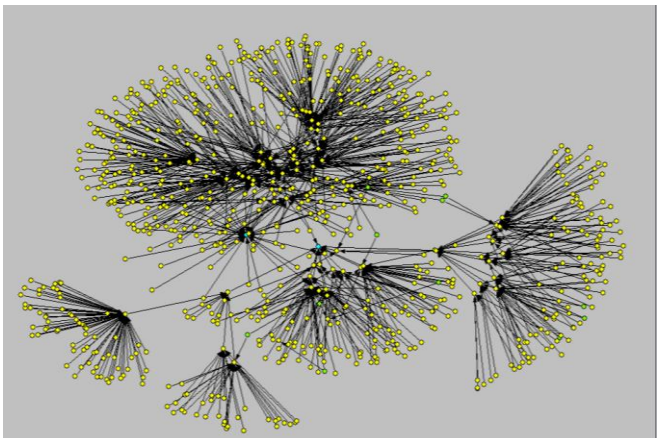

Fig. 1 Pajek output with Kamada Kawi free algorithm

When number of records is increased to 13000, the resultant graph of appying the same Kamada Kawi alogorithm is like the following figure.

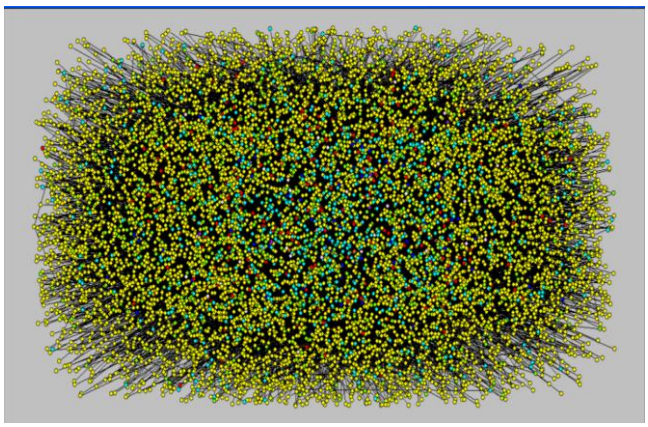

Fig. 2 Pajek output with Kamada Kawi free algorithm

Getting cited-by papers information of even a single paper is not possible at all from this graph. The same is the case with other layouts when the number of records is more. With Fruchterman reingold 3D algorithm, the result with 13000 citation records is about same as shown in the following figure.

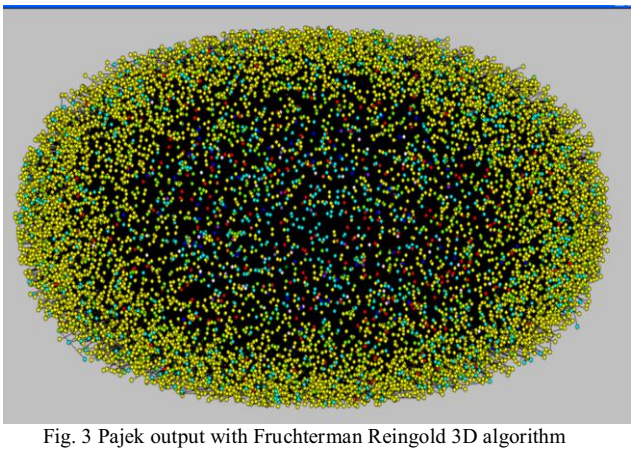

Fig. 3 Pajek output with Fruchterman Reingold 3D algorithm

We tried with different layouts even with other tools like in GUESS we used spring layout (Pajek has not the support of this layout) with 1600 citation records that was giving better visualization than Pajek layouts as shown in Fig. 5.

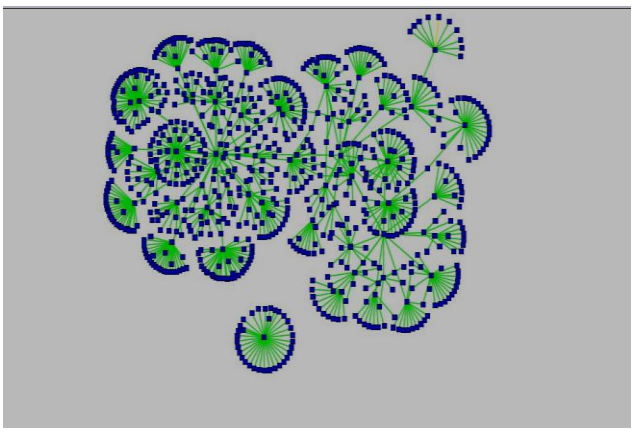

Fig. 4 GUESS layout with spring layout

Here still tracing the paths from one node to another in order to find out cited-by papers is very cumbersome. Also these visualizations are static images that can only be zoomed in and out in their own tool windows with special key combinations. To make the Pajek outputs dynamic, we used JUNG libraries in Java in order to show the specific citation hierarchies of any node. GUESS graphs can also be made dynamic using Python language but we used Java with Pajek graphs as this work is the part of overall project to show the nature of relationships between the research papers that we are implementing in Java.
In order to provide a comprehensive visualization of cited-by papers, we attached the mouse move event to pajek graph nodes. When mouse pointer comes over a node; the hierarchy of cited-by papers starting from that node is visualized in the other pane. For paper-id 75, the cited-by papers are visualized like following diagram.

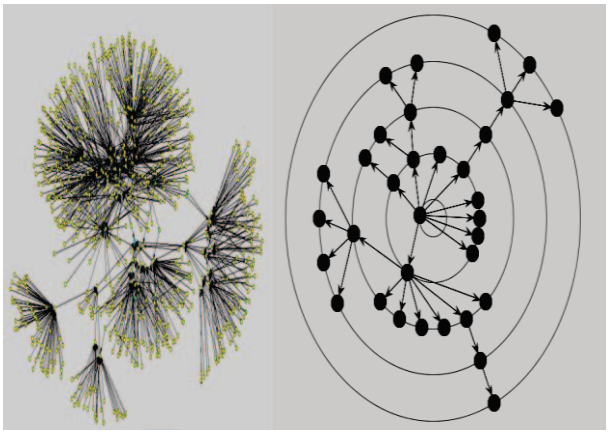

Fig. 5 Java result

If the hierarchy of cited-by papers increases then zoom-in still hides the nodes. In order to keep all the information visible, we have provided an option of a lens that is resizable and inside it the required part can be magnified. The position of the graph can be changed in it as well according to user requirement that gives the hyperbolic visualization as shown below.

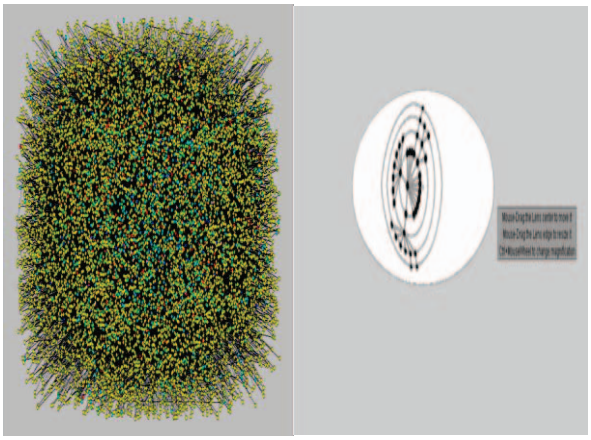

Fig. 6 java result

V. CONCLUSION AND FUTURE WORK

In this paper we showed that the graph of citation data in any layout doesn't provide the quick and clear information about the cited-by papers of some focus paper due to the nature of data. As the graph grows, its density increases and its visualization keep on decreasing to show the cited-by papers until it becomes useless. We extract only the information about cited-by papers of a particular paper from this large graph and in response to mouse move event of the node, reflect it in the other frame. Visualization used to show this sub graph is radial tree with lens for hyperbolic view and magnification purpose which is simple, and easy to understand.

As this paper is part of the overall project of finding the citation functions, so in future, we will visualize the citation functions by using different colors for the nodes. We also plan to arrange the small sub graphs that represent particular paper citations, in a radial tree form according to the years of publications. Circles will be representing the years, so all cited-by papers of a particular year will be on one circle.

REFERENCES

[1] Garfield, E., "Citation Indexing—Its Theory and Application in Science, Technology, and Humanities," *New York: John Wiley*, 1979.

[2] Garfield, E., Sher, I.H., & Torpie, R.J., " The Use of Citation Data in Writing the History of Science," *Philadelphia: Institute for Scientific Information*, 1964.

[3] Hummon, N.P. & Doreian, P.," Connectivity in a Citation Network: The Development of DNA Theory". Social Networks, 11, 39–63, 1989

[4] Baldi, S. & Hargens, L.L.. "Reference Network Structure in turn-of-the-Century Physics: The Case of N-rays,". In *M.E.D. Koenig & A. Bookstein (Eds.), Proceedings of the Fifth Biennial Conference of the International Society for Scientometrics and Informetrics* (pp. 43–52), Medford, NJ: Learned Information, 1995.

[5] Yermish, I.. "A Citation Based Interactive Associative Information Retrieval System". Unpublished doctoral dissertation, University of Pennsylvania, Philadelphia, PA, 1975.

[6] Oppenheim, C., & Renn, S.P., "Highly Cited Old Papers and the Reasons Why They Continue to be Cited,". *Journal of the American Society for Information Science*, 29, 225–231, 1978.

[7] Peter J. Denning. "The ACM Digital Library Goes Live". Communications of the ACM, 40(7):28–29, July 1997.

[8] http://jung.sourceforge.net/

[9] Borgatti S, Everett M, Freeman L. "UCINET: Software for Social Network Analysis," 2004, http://www.analytictech.com/ucinet

[10] Batagelj V, Mrvar A. "Pajek: Program for Large Network Analysis," 2004 http://vlado.fmf.uni-lj.si/pub/networks/pajek/

[11] R Development Core Team. R: "A Language and Environment for Statistical Computing," 2004. *R Foundation for Statistical Computing, Vienna, Austria*. 3-900051-07-0, URL http:// www.R-project.org

[12] Butts C. "S Routines for Social Network Analysis in the R Environment: The SNA Package. 2004" http://erzuli.ss.uci.edu/R.stuff/.

[13] IBM Corporation (1999). "GFC: Graph Foundation Classes for Java." http://www.alphaworks.ibm.com/tech/gfc

[14] Adar, Eytan, "GUESS: A Language and Interface for Graph Exploration," CHI 2006 http://graphexploration.cond.org/documentation.html

[15] Cyram Company, Ltd. "NetMiner," 2004 http://www.netminer.com. Version 2.5.

[16] Boer P, Huisman M, Snijders TAB, Zeggelink EP. "StOCNET: an Open Software System for the Advanced Statistical Analysis of Social Networks." 2003 http://stat.gamma.rug.nl/stocnet Version 1.4. Gronigen: ProGAMMA / ICS.

[17] Richards W, Seary A (2004). http://www.sfu.ca/~richards/Multinet/Pages/multinet.htm. Version 4.55.

[18] Fekete JD, "The InfoViz Toolkit." In "Proceedings of the 10th IEEE Symposium on Information Visualization (InfoVis'04)," 2004, pp. 167–174. IEEE Press.

[19] Penumarthy S, Mane KK, B¨orner K. "InfoVis Cyber Infrastructure Software Framework.", 2004 http://iv.slis.indiana.edu/sw/

[20] Brandes U, Wagner D. "visone - Analysis and Visualization of Social Networks," 2003, In M J¨unger, P Mutzel (eds.), "Graph Drawing Software," pp. 321–340. Springer-Verlag.

[21] Siek JG, Lee LQ, Lumsdaine A. "The Boost Graph Library User Guide and Reference Manual. C++," 2001, In-Depth Series. Addison-Wesley.

[22] Alder G. "JGraph: The Java Graph Visualization Library." 2005, http://www.jgraph.com

[23] yWorks (2004). "yFiles." http://www.yworks.com/en/products_yfiles_about.htm

[24] C. Lee Giles, Kurt Bollacker, and Steve Lawrence. "CiteSeer: An Automatic Citation Indexing System," In Digital Libraries 98 - The Third ACM Conference on Digital Libraries, pages 89–98, June 1998.