IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 4, April 2016

A Graph Data Summarization and Data Visualization Technique

Stibu Stephen[1], Dr. R. Manicka Chezian[2]

PhD. Scholar, Department of Computer Science, N.G.M College, Pollachi, Tamil Nadu, India[1]

Associate Professor, Department of Computer Science, N.G.M College, Pollachi, Tamil Nadu, India[2]

Abstract: A data Analysis and visualization technique poses challenges too many areas of the data mining research. There are several visualization techniques and tools have been proposed for almost all domains. But there is a necessity to summarize and visualize a large citation network data according to the user customization. While performing the visualization, influence data should be identified using the summarization technique. The summarization and visualization of graph structured data is a tedious part in research. The existing state of the art influence maximization algorithms can detect the most influential node in a citation network for all structured data, except graph structured data. Clustering techniques are widely used to fold large graph structured data in existing graphs summarization methods. In this paper, first formally define the problem of data summarization and visualization process with three segments, which are effective summarization, localized summarization and handling high, influenced rich information in citation networks. In general, research filed contains lots of interrelated datas, which has multi associations among the data. To handle the above Graph data visualization and summarization problem, here propose a new prototype named as (GSV) Graph Summarization and GSV algorithm for large scale citation networks. Finally present a theoretical analysis on GSV, which is equivalent to the existing kernel k mean clustering algorithm.

Keywords: Data summarization, visual data mining, Graph mining, GSV.

I. INTRODUCTION

The exponential growth of hardware and software technologies allows system to store huge amount of data. Usually there is an enormous space from the stored data to the knowledge that could be interpreted from the data. In basic Data Analysis, only some initial knowledge is known about the data, but Data Mining could help in a more deep knowledge about the data [1]. Seeking knowledge from vast data is one of the most desired attributes of Data Mining. Data analysis by manual is only suitable for tiny dataset, not for large dataset like big data. So data mining involves the use of sophisticated data analysis tools to analyze huge data and discover previously unknown data's, valid patterns and relationships. Those tools developed with various techniques and features to satisfy the new demands in complex data analysis. The data analysis and Visualization has high demand in several applications such as finance, education etc., which helps to acquire visual representation of structured and unstructured data [2]. In more specific, user interactive applications are more inspired by the data analysis and visualization techniques. These techniques are segmented into several branches, such as displaying summarized properties of data, investigating huge database and exploring various relationships between several information's and finally this has the analysis part of geographical and spatial domains too. Data analysis and visualization is applied in almost all domains [3]. In information visualization and summarization, the data usually consists of a large number of raw information's, each consisting of a number of variables or aspects. For example e-commerce transactions, public census information's etc., The above stated Data sets may be one-dimensional, two-dimensional or multi-dimensional and may have more complex data types such as text or hypertext or hierarchies or graphs.

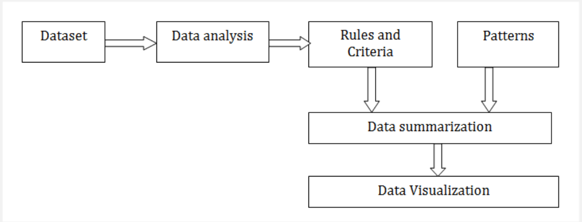


Fig: 1.0 process of data analysis and visualization

The fig 1.0 represents the basic process of data analysis and visualization. This includes the various steps such as data summarization and rule specification. Data records in Citation networks have more dependencies with other information's. Graphs are widely used to represent such interdependencies [4]. A graph consists of set of objects, called nodes and connections between these objects, called edges. Examples are the e-mail interrelationships among people, their shopping behavior, and the file structure of the hard disk or the hyperlinks in the World Wide Web. There are a number of specific visualization techniques that deal with hierarchical and graphical data. A nice overview of hierarchical information visualization techniques can be found in [5], an overview of web visualization techniques at [6] and an overview book on all aspects related to graph drawing is [7]. In this paper explore the contribution of data visualization technique in books citation domain.

DOI 10.17148/IJARCCE.2016.5472

II. LITERATURE REVIEW

In modern research activities, CITATION networks [8] are more requisite to understand fundamental activities of many research aspects. This is a fundamental resource for the data mining field, which analyze the relationship of those publication activities such as research topics, communities and trends etc. the way of summarizing citation graphs, which establishes the above relationships in a graphical way is referred as Influence Graph Summarization problem [9]. This chapter reviews the related work from various aspects. The first one is summarization and visualization techniques. The next one is graph summarization and simplification. First, data visualization techniques can be evaluated using two main groups, which are interface and features of dataset. The visualization tool needs to be able to deal with this overlap. Several visualization tools have been developed. All the visualization techniques and tools have the ability to play a vital role in exploratory visualization. The visualization can be integrated with other data mining tools in the field of knowledge discovery process. The Growing Self Organizing Map (GSOM) [10] is an approach to use neural networks, which supports unsupervised learning. The GSOM has a presentation graph, so this is called as Feature Map. This type of GSOM provides better presentation visualizations compared to the previous feature maps from a simple Self Organizing Map (SOM). Thus the exploratory multi-dimensional visualization tool can be used to further automate the data analysis/mining using a GSOM. These techniques are integrated with other data mining techniques and the tool as named as DBMiner, which integrates data mining approaches with visualization components.

Second, the graph summarization and simplifications are discussed. In visualization and summarization, various types of data and tools were used. Constructing minor summarizations to represent a outsized graph has been a challenging task, especially using graph clustering algorithms [11]. Several interesting work has been done in the context of graph clustering and summarization. The graph summarization methods based on node attributes ensure the content coherence on clusters, but again they are not tailored for the flow rate maximization objective in the IGS problem. These works typically look for coherent regions in the graph by optimizing a predefined loss function, which minimizes the inter-cluster connections and maximizing the intra cluster attribute homogeneity. But some techniques fail to reveal the influence flows important for the IGS problem. The social graph simplification performs the information dispersal over social networks such as Facebook and Twitter; several researchers have studied the problem of extracting the most important social paths based on information propagation logs to optimize applications such as viral marketing. Finally, decision making from the above summarized and visualized data will only complete, if the detection data has more influence. Due to the huge size of data, influence maximization is mandatory.

III. PROBLEM DEFINITION

Effective Visualization
The primary goal of GS is to summarize the important data flows from a source node in the reversed citation graph. For the effective visualization, it is defined as the objective of maximizing the overall flow rate given the number of flows to display. The consistency within the generated node cluster is not defined by the dense internal connection any more, but rather by the high node topological similarity in the same cluster. In this objective, more edges will be cut across clusters than traditional methods, so as to highlight the inter-cluster flows that outline the influence patterns.

Localized summarization:
While a full citation graph can span millions of nodes and prohibit any readable visual summarization, in the GS objective, here is switch to summarize the influence of a single source node. This localized summarization problem is at least as important as the global summarization. Consider a user navigating the influence graph of computer science papers, after an overview of the entire field, most likely she will drill down to a few key papers and examine their influences separately.

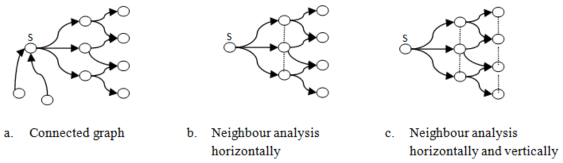
Rich information:
The citation graphs have rich node attributes such as the venue, paper topic and research topic and often evolve over time known as the publication date. Incorporating this information to optimize the summarization result poses additional challenges to our work.

IV. PROPOSED SYSTEM

To solve the GS problem, here propose an algorithm over which, build a prototype system called GSV, to generate hierarchical, Graph Summarization and Visualization over large scale citation networks. The algorithm GSV is flexible and confesses many existing graph mining algorithms.
The main contributions of this paper can be summarized as:

- Analyses the Graph summarization related issues
- Our proposal finds a new algorithm to perform graph visualization in large scale citation networks.
- Finds the maximum influenced data from the multi dimensional dataset.

GSV algorithm:
To solve the GS problem, here propose an algorithm GSV, which performs data summarization and visualization, as illustrated in Fig. 2.0.



a. Connected graph
b. Neighbour analysis horizontally
c. Neighbour analysis horizontally and vertically

Fig 2.0 GSV Graph summarization working process

From the graph shown in summarization illustration of fig 2.0, the initial graph structured data has taken as an input, where the graph started with node S. the source node considered as an author. Based on the first node information's, the neighbour analysis will be performed. For this process, the basic kernel K-means algorithm is used. The initial clustering with k neighbour results in horizontal format gives the initial summary state and the final step c represents the vertical clustering after the successful horizontal graph construction.

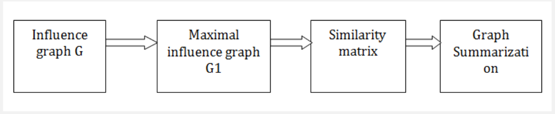


Fig 3.0 GSV overall working process

In the citation networks, the source node often carries some additional information, such as the research topic, author details and the venue along with the publication date of a specific paper,. This information, beyond the network topology, can be critical in many scenarios. So the above fig 3.0 shows the overall working process of GSV in the above stated domain. This performs the similarity matrix in every stage and performs the visualization with summarized data. The following steps explain the above GSV process in detail.

Algorithm: GSV algorithm
Input: Graph structured data with initial source node
Output: summarized data visualization

Steps:
Step 1: Get influenced graph G
Step 2: Pick source node S
Step 3: Rooted search on G from S
Step 4: find maximal influence graph G1

a) Compute topology similarity matrix M^{G1}
b) Node attribute adjacency matrix A^{D}
c) Generalized similarity matrix M^{D}-considering node attributes

Step 5: Specify I,J, where I is the number of clusters and J is the flows
Step 6: compute node summarization by matrix decomposition
Step 7: Post process of link and prune irrelevant links.

Fig 4.0 GSV Algorithm

Initially, the maximal influence graph G1 is computed from the input graph G by a breadth-first or depth-first search starting from the source node S. Over the maximal influence graph G1, a few processing components work in parallel to generate several matrices from the graph: the topology similarity matrix MG1, the optional node attribute adjacency matrix AD and the generalized similarity matrix MD. The core of the algorithm GSV is the decomposition of the similarity matrices to generate I node clusters for the summarization. Here carefully design the topology similarity matrix to ensure that the graph summarization approximates the flow rate maximization objective. The optional node attribute adjacency matrices can be incorporated to ensure coherence on node attributes while still optimizing the proposed objective.

V. CONCLUSION

In this paper consider the problem of summarizing influences in large citation networks under a flow-based and localized context. Formally define this as an optimization problem, study its linkage to the existing clustering methods, and present an algorithm as well as the prototype system GSV to solve it. GSV achieves all the three design objectives, including: (1) data summarization and analysis; (2) a localized visual summarization from the source node; and (3) graph visualization. Here describe both the matrix decomposition based main algorithm and the implementation details of GSV. Through comprehensive evaluations with real-world citation networks, Demonstrate that the proposed algorithm constantly outperforms classical methods, such as the graph summarization and visualization algorithms.

REFERENCES

[1]. A. Buja, D. F. Swayne, and D. Cook, “Interactive high dimensional data visualization,” Journal of Computational and Graphical Statistics, vol. 5, no. 1, pp. 78–99, 1996.
[2]. G. Jeh and J. Widom, “SimRank: A measure of Structural-context similarity,” in Proc. 8th ACM SIGKDD Int. Conf. Knowledge. Discovery Data Mining, pp. 538–543., 2002
[3]. B. Spence, Information Visualization, Pearson Education Higher Education publishers, UK, 2000.
[4]. Shi, Lei, et al. "VEGAS: Visual influence Graph Summarization on Citation Networks." Knowledge and Data Engineering, IEEE Transactions on 27.12 (2015)
[5]. C. Chen, Information Visualization and Virtual Environments, Springer-Verlag, London, 1999.
[6]. M. Dodge, “Web visualization,” http://www.geog.ucl.ac.uk/ casa/martin/geography of cyberspace.html, Oct 2001.
[7]. G. D. Battista, P. Eades, R. Tamassia, and I. G. Tollis, Graph Drawing, Prentice Hall, 1999.
[8]. J. E. Hirsch, “An index to quantify an individual's scientific research output,” Proc. Nat. Acad. Sci. United States of America, vol. 102, no. 46, pp. 16 569–16 572, 2005.
[9]. Y. Tian, R. A. Hankins, and J. M. Patel, “Efficient aggregation for graph summarization,” in Proc. ACM SIGMOD Int. Conf. Manag. Data, pp. 567–580., 2008
[10]. Alahakoon D., Halgamuge S.K. and Srinivasan B.; Dynamic Self Organizing Maps with Controlled Growth for Knowledge Discovery, IEEE Transactions on Neural Networks (Special Issue on Data Mining), Volume 11, No. 3, pp 601-614., May 2000
[11]. S. Fortunato, “Community detection in graphs,” Phys. Rep., vol. 486, no. 3–5, pp. 75–174, 2010.
[12]. Y. Tian, R. A. Hankins, and J. M. Patel, “Efficient aggregation for graph summarization,” in Proc. ACM SIGMOD International. Configuration. Management. Data, pp. 567–580, 2008.
[13]. N. Zhang, Y. Tian, and J. M. Patel, “Discovery-driven graph summarization,” in Proc. IEEE 26th Int. Conf. Data Eng. pp. 880–891, 2010,.