International Journal of Engineering, Management, Humanities and Social Sciences Paradigms (IJEMHS)
(Volume 28, Issue 03) Publishing Month: August 2017
An Indexed and Referred Journal with Impact Factor: 2.75
ISSN: 2347-601X
www.ijemhs.com

Examining Subtree and Subgraph Mining Techniques

[1]Arpita Swain
Gandhi Institute of Excellent Technocrats, Bhubaneswar, India
[2]Samrat Ranjan Nayak
Ghanashyam Hemalata Institute of Technology and Management, Puri, Odisha, India

Abstract
Graph is a basic data structure which, can be used to model complex structures and the relationships between the data elements, such as XML documents, social networks, communication networks, chemical informatics, biology networks, and structure of web pages. In graph mining scenario Frequent subgraph pattern mining is one of the important aspect. Researchers have found may research oriented topic in this area, such as analysis and processing of XML documents, documents clustering and classification, images and video indexing, graph indexing for graph querying, routing in computer networks, web links analysis, drugs design, and carcinogenesis. Several frequent mining algorithms use various methods on different datasets, patterns mining types, graph and tree representations.This paper presents a brief report of an intensive investigation of frequent subgraphs and subtrees mining algorithms.

Keywords
Graph Mining, Subgraph, Frequent Pattern, Graph indexing.

I. INTRODUCTION

Today we are faced with ever-increasing volumes of data. Most of these data naturally are of graph or tree structure. The process of extracting new and useful knowledge from graph data is known as graph mining [1] [2] Frequent subgraph patterns mining [3] is an important part of graph mining. It is defined as "process of pattern extraction from a database that the number frequency of which is greater than or equal to a threshold defined by the user." Due to its wide utilization in various fields, including social networkanalysis[4][5][6],XMLdocumentsclusteringandclassification
[7] [8], network intrusion [9] [10], VLSI reverse [11], behavioral modeling [12], semantic web [13], graph indexing [14] [15] [16] [17] [18], web logs analysis[19], links analysis[20], drug design [21] [22] [23], and Classification of chemical compounds[24] [25] [26], this field has been subject matter of several works.

The present paper is an attempt to survey subtree and subgraph mining algorithms. A comparison and classification of these algorithms, according to their different features, is also made. The next section discusses the literature review followed by section three that deals with the basic ideas and concepts of graphs and trees. Mining algorithms, frequent subgraphs are discussed in section four from different viewpoint such as criteria of representing graphs (adjacency matrix and adjacency list), generation of subgraphs, number of replications, pattern growth-based and apriori-based classifications, classification based on search method, classification based on transactional and single inputs, classification based on type of output, and also Mining based on the logic. Fifth section focuses on frequent Mining algorithm from different angles such as trees representation method, type of algorithms input, tree-based Mining, and Mining based on Constraints on outputs.

II. RELATEDWORKS

H.J.Patel1, R.Prajapati,et al. [27] Classified graph mining and mentioned two types of the algorithms, apriori-based and pattern growth based. K.Lakshmi1,T.Meyyappan [28] studied apriori based and pattern growth based, taking into account aspects such as input/output type, how to display a graph, how to generate candidates, and

how many times a candidates is repeated in the graph dataset. In [29] D.Kavitha, B.V.Manikyala, et al. suggested the third type of graph mining algorithms named as inductive logic programming. Here a complete survey of graph mining concepts and a very useful set of examples to ease the understanding of the concept come next.

BASICCONCEPTS

Garph

A graph G (V, E) is composed of a set of vertices (V) connected to each other by and a set of edges (E).

Tree

A tree T is a connected graph that has no cycle. In other words, there is only and only one path between any two vertices.

Subgraph

A **subgraph** G '(V', E') is a subgraph of G (V, E), which vertices and edges are subsets of V and E respectively:

- $V' \subseteq V$

One may say that a subgraph of a graph is a pattern of that graph. Concerning trees two types of patterns can be defined:

Induced pattern

The definition is exactly the same as the definition of subtree in a tree (Figure.1.a, Figure.1.c). It means that the vertices and the edges of Figure.1.a. Can be seen in Figure.1.c as well

Embeddedpattern:

Almost the same as induced pattern, except that there may be one or more supplementary vertices between the two parents and child nodes of pattern, For example vertex **A** in Figure.1.c is parent of vertex D; and in Figure.1.b an embedded pattern of Figure 1.c is seen.

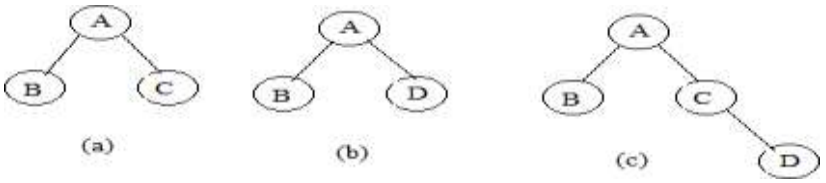

Figure.1. An example of the Induced and embedded subtree pattern

Isomorphism

Two graphs are **isomorph**, if there are one to one relationships among their vertices and edges.

FrequentSubgraph

Suppose a graph G and a set of graphs D = {g1, g2, g3,…, gn} are given, support(G) is:

$$Support\ (G) = \frac{\quad}{\quad}$$

A graph G in a dataset D is called **Frequent** if its support is not less than of a predefined threshold.

III. CONCLUSION

Frequent subgraph Mining algorithms were first examined from different viewpoints such as different ways of representing a graph (e.g. adjacency matrix and adjacency list), generation of subgraphs, frequency counting, pattern growth-based and apriori-based algorithm classification, search based classification, input-based classification (single, transactional), output based classification. Furthermore, Mining based on logic was discussed. Afterward, frequent subtrees traversal algorithms were examined from different viewpoints such as trees representation methods, type of inputs, tree-based traversal, and also Mining based on Constraints of outputs. Given theresults,itisconcludedthatinabsenceofgeneratingpatternsbypattern-growth, it is featured

with less computation work and needs smaller memory size. Moreover, these algorithms are specifically designed for trees and graphs and cannot be used for other purposes. On the other hand, as they work on variety of datasets, it is not easy to find tradeoffs between them. The same frequent patterns can be used for searching similarity, indexing, classifying graphs and documents in future studies. Parallel methods and technologies such as Hadoop can also be needed when working with excessive datavolume.

IV. REFERENCES

[1] A.Rajaraman, J.D.Ullman, 2012. *Mining of Massive Datasets*, 2nded.

[2] J.Han, M.Kamber, 2006, *Data Mining Concepts and Techniques*. USA: Diane Cerra.

[3] Kuramochi, Michihiro, and G.Karypis., 2004. An efficient algorithm for discovering frequent subgraphs, in *IEEE Transactions on Knowledge and Data Engineering*, pp.1038-1051.

[4] J.Huan, W.Wang, J. Prins, 2003. Efficient Mining of Frequent Subgraph in the presence of isomorphism, in *Third IEEE International Conference on Data Minign(ICDM)*.

[5] (2013, Dec.) Trust Network Datasets - TrustLet. [Online].http://www.trustlet.org

[6] L.YAN, J.WANG, 2011. Extracting regular behaviors from social media networks, in *Third International Conference on Multimedia Information Networking andSecurity*.

[7] Ivancsy,I. Renata, I.Vajk., 2009. Clustering XML documents using frequent subtrees, *Advances in Focused Retrieval*, Vol. 3, pp.436-445.

[8] J.Yuan, X.Li, L.Ma, 2008. An Improved XML Document Clustering Using Path Features, in *Fifth International Conference on Fuzzy Systems and knowledge Discovery*, Vol.2.

[9] Lee, Wenke, and Salvatore J. Stolfo, 2000. A framework for constructing features and models for intrusion detection systems, in *ACM transactions on Information and system security (TiSSEC)*, pp.227-261.

[10] Ko, C, Logic induction of valid behavior specifications for intrusiondetection, 2000. in *In IEEE Symposium on Security and Privacy (S&P)*, pp. 142–155.

[11] Yoshida, K. and Motoda, 1995. CLIP: Concept learning from inference patterns, in *Artificial Intelligence*, pp.63–92.

[12] Wasserman, S., Faust, K., and Iacobucci. D, 1994. *Social network analysis : Methods and applications*. Cambridge universityPress.

[13] Berendt, B., Hotho, A., and Stumme, G., 2002. semantic web mining, in *In Conference International Semantic Web (ISWC)*, pp.264–278.

[14] S.C.Manekar, M.Narnaware, May 2013. Indexing Frequent Subgraphs in Large graph Database using Parallelization, *International Journal of Science and Research (IJSR)*, Vol. 2 , No.5.

[15] Peng, Tao, et al., 2010. A Graph Indexing Approach for Content-Based Recommendation System, in *IEEE Second International Conference on Multimedia and Information Technology (MMIT)*, pp.93-97.

[16] S.Sakr, E.Pardede, 2011. Graph Data Management: Techniques and Applications, in *Published in the United States of America by Information ScienceReference*.

[17] Y.Xiaogang, T.Ye, P.Tao, C.Canfeng, M.Jian, 2010. Semantic-Based Graph Index for Mobile Photo Search," in *Second International Workshop on Education Technology and Computer Science*, pp.193-197.

[18] Yildirim, Hilmi, and Mohammed Javeed Zaki., 2010. Graph indexing for reachability queries, in *26th International Conference on Data Engineering Workshops (ICDEW)IEEE*, pp. 321-324.

[19] R.IvancsyandI.Vajk,2006.FrequentPatternMininginWebLogData,in *Acta PolytechnicaHungarica*, pp. 77-90.

[20] G.XU, Y.zhang, L.li, 2010. *Web mining and Social Networking*. melbourn: Springer.

[21] S.Ranu, A.K. Singh, 2010. Indexing and mining topological patterns for drug, in *ACM, Data mining and knowlodge discovery*, Berlin,Germany.

[22] (2013, Dec.) Drug Information Portal. [Online].http://druginfo.nlm.nih.gov

[23] (2013, Dec.) DrugBank. [Online].http://www.drugbank.ca

[24] Dehaspe,Toivonen, and King, R.D., 1998. Finding frequent substructures in chemical compounds, in *In Proc. of the 4th ACM International Conference on Knowledge Discovery and Data Mining*,pp.30-36.

[25] Kramer, S., De Raedt, L., and Helma, C., 2001. Molecular feature mining in HIV data, in *In Proc. of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-01)*, pp.136–