An Iterative Mapreduce based Frequent Subgraph Mining Algorithm with Load Balancing

Ms. Vishnuppriya S (1), Sandeepkumar R (2), Swathi S (3), Vishnu Priya T. S (4)
Assistant Professor (1)
Final Year(2),(3),(4)
Department of CSE
Velalar College of Engineering and Technology Thindal, Erode

***Abstract*:- Health plays a major role in human's life. Healthcare insurance is provided by insurance company in order to reduce the financial costs. Nowadays, frauds in healthcare insurance was occurring and it has caused huge dollar losses all over the world. It is mainly due to the disease that occurs for more than three months which is called as chronic disease. If the investigators understand the evolution of disease earlier means they can save their insurance amount by detecting the insurance frauds earlier. In the existing method, Frequent Subgraph Mining Algorithm is used. Firstly, a graph for each patient is constructed. To mine the subgraph from that graph, Frequent Subgraph Mining is used. By using that subgraph the base disease is found. If the base disease and chronic rules satisfies means the insurance is claimed. But, it won't solve all type of problems. That means for the same chronic disease, different medication stages is not considered. All the process is done in single system whatever the system capability is and the overall time efficiency is less. In the proposed method, Load Balancing Algorithm is used along with the Frequent Subgraph Mining Algorithm. Before the nodes are assigned for mapper process, they are equally balanced by using Load Balancing. That is in the sub graph, small graphs are assigned to node A and one big graph is assigned to node B. Here for the same chronic disease, different medication stages is considered. The time efficiency is increased by balancing the process of the system.**

Keywords: Evolution of disease; subgraph mining; chronic disease

1. INTRODUCTION:

Data mining is extracting information in the huge amount of data. Hidden information from the large datasets is discovered. It is a very new and useful technology that helps companies to focus on information which is in the data warehouses. Tools of data mining helps to predict the future trends and allows to make decisions in the business. It answers all the business questions that traditionally takes too time to resolve. Then, it also helps organization to make use of the data stored in their databases. When it comes to decision making, it is true in different types of fields and organizations. Knowledge Discovery in Databases (KDD) is defined as the process of discovering useful information from the given data's collection. The techniques of data mining is a process that includes Data cleaning, Data integration, Data selection Data transformation, Data mining, Pattern evaluation, Knowledge representation like that.

In the process of data cleaning the data's that are not relevant are removed and the data's that are common will be available in the integration. In data selection the data's that are relevant is selected and available data is transformed to the procedure of mining in transformation. The next phase is Data mining which means extracting the patterns. According to the measures, knowledge from the available data is identified in pattern evaluation. The final step in the process is the knowledge discovery which here means as the discovering of knowledge in the data mining.

Data mining techniques is used to help the users and to understand and to interpret the data mining results. These methods are applied to uncover the hidden patterns and it has been used for many years. The main reason for using data mining is to observe and analyze the behavior. The fact is that the data subsets analyzed may or may not belong to the whole domain. Some categories are involved in the function of data mining.

The tasks of data mining are classified as Predictive and Descriptive. Predictive data mining means prediction about future data sets. Descriptive data mining means it will describe new or latest information. The data mining method is used in many of the sectors. For example, the chronic disease will be identified and it also will track the regions about the spreading of disease and some programs are designed to reduce those spreading of disease. The professionals of healthcare will analyze about the diseases along with the regions of patients with maximum admissions in the hospital. With this information, they will make a way for giving awareness to the people. By giving the awareness, the patient count will be reduced. This will reduce the number of patients admitted in the hospital.

2. BACKGROUND:

2.1 AUTHORS : Bay Vo and Bac Le

In this paper, we present a brand new set of rules for mining generalized affiliation policies. We expand the set of rules which scans the database one time simplest and use Tidset to compute the support generalized itemset faster. A tree structure is present here which is known as GIT-tree, an extension of IT-tree, is an advanced to shop database for data mining that is the common itemsets from the type of hierarchical form of database. Our set of rules very faster than this MMS_Cumulate,an algorithm mining a frequent itemsets in the hierarchical type of the database within those of the couple of minimum support provided in the given set of the experimental databases.

Mining association policies will always play an very important function roles in such type of the expertise type of discover and data mining that is known as (KDD). Its motive is mining the hidden know-how in the databases that are available. Mining affiliation guidelines in the type hierarchical database has been proposed. Mining association rules amongst gadget in the type of the hierarchical tree that satisfy minSup and the minConf. However, some of the study does now not change the support in exceptional hierarchical levels.

The paper also proposed the uniform minimum assist in each and also in every stage, the gadgets that are present in the identical degree will always balanced and get the equal minimal support. Hence, mining association rules among itemsets in the equal level. The other rule is Association rule which is used as an technique of the mining helps which becomes proposed. This permits customers identify the exclusive minimum helps for the rare items, so we can mine both common and rare guidelines. However, considering and evaluating all the techniques used here will does not get traverse so the whole hierarchism now and always so that it is very difficult to locate association guidelines among items in one of kind levels.

2.2 AUTHORS : Bahman Bahmani, Ravi Kumar, Sergei Vassilvitskii

The problem of locating domestically dense additives of a graph is an critical and complex primitive in the records analysis with an wide-ranging applications. In this paper, we just gift the new algorithms for finding the subgraph in the streaming model that is present. For any > 0, our algorithms make O(log1+n) passes over the enter and to find a subgraph whose is assured to be inside a component 2(1 +) of the optimum. Here, Our algorithms are also easily parallelizable and we illustrate this by knowing them inside the MapReduce model.

Also, here the problem of finding dense subgraphs, in the numerous data control packages, in streaming and also in the MapReduce, two computational fashions which is an increasing number of being adopted by the means of the large-scale of processing applications.We showed a simple set of rules that make a small number of passes over the graph and obtains a (2+)-approximation to the densest subgraph.

For the case while the subgraph is more than a positive size and then only the graph is directed. Those are the subgraph which sincerely scale but over the guarantees. Then, here it comprises of some of our experiments confirmed surely that the algorithms are indeed scalable here and achieve great and overall the performance frequency will much higher than theoretical guarantees. Our set of rules scalability is the main motive it was feasible to run it on a graph with more than a half of one of the billion nodes and six billion edges.

2.3 AUTHORS : Jie Tang, Jimeng Sun, Chi Wang and Zi Yang

We examine a novel trouble of subject matter-based which is based totally on the social evaluation and influence. So, that we are recommend a approach Topical Affinity Propagation (TAP) which is just to be used to explain about the trouble that is the usage of a graphical type of the probabilistic version. To deal these kind of the efficient problem, we want just want to present a here a very latest and brand new algorithm for that those type of the TFG version. A distributed learning set of rules has been implemented.

Then further, it will also give some of the result as per experimental effects on three the different type of the data sets which here exhibit that the method which was proposed can efficiently find out the subject based on the social influences. The distributed getting to know set of rules additionally has a very good scalability and also the overall performance. We practice the proposed technique to finding. The Experiments will show that the determined subject is of matter-based on the influences by the means of the proposed method and it will always improve the performance of expert finding.

General trouble of community influence evaluation represents a very brand new and interesting research path in the mining type which is social mining. There are so many instructions is here in this work. Interesting trouble is here to extend the TFG version.

Also, here exist an another complex type of issue which layout the TAP technique for the semi supervised studying. Here, the Users may or may not offer feedbacks for evaluation.

2.4 AUTHORS : U Kang, Charalampos E. Tsourakakis, Christos Faloutsos

We describe a very important primitive for the PEGASUS which is known as the GIM-V. That is an very important and useful one. GIM-V, Which is known as Generalized Iterated Matrix-Vector multiplication. This is particularly optimized and achieving the (a) desirable scale-up at those range of the given available machines (b)The linear running time at the variety of edges (c) greater than 5 times faster performance over the non-optimized version of GIM-V. Our experiments ran on M45, certainly one of the pinnacle 50 which will be considered as supercomputers in the world.

Here, our findings on the document of several real graphs, including one of the biggest publicly available of the type of Web Graphs, thanks to yahoo. The Graphs are very ubiquitous here: pc network also the cell call networks and www [1], protein regulation networks to call a few. The huge extent will only will have the data, stunning fulfillment of the type of the on-line social networks.

In Web2.0 applications all lead to the graphs of exceptional size. Then, the typical graph mining algorithms silently anticipate that the available graph will fit in reminiscence of a regular workstation, or it will be atleast on a unmarried disk; the above graphs just violate above assumptions and it span more than the one Giga-bytes, and heading to Tera and Peta-bytes of data.

3. EXISTING METHODS :

The existing system proposes a method to mine the chronic disease progression and helps us to detect chronic disease- related healthcare insurance fraud.

The steps are mentioned below:

Step 1 : Construct a health seeking temporal graph for each patient.

Step 2 : From the frequent disease-process subgraphs, Constrained Frequent Subgraph Mining (CFSM) algorithm is used and then the subgraph is constructed.

Step 3 : From the subgraph, Construct the base disease progression network by aggregation of the recoded graph.

Step 4 : Conduct community detection on the base disease progression network and transform them into chronic disease progression rules.

Step 5:Conduct chronic disease-based healthcare insurance fraud detection according to the rules obtained in Step 4 Even if the number of graphs is more, the application executes. That is, mine them in single system.

3.1 DRAWBACKS :

- All the graphs are processed in single system even if the system capability is less.
- For example, one node may be assigned with more big graphs and other node with small graphs.
- Overall time efficiency is poor.

4. PROPOSED SYSTEM :

Construct a health seeking temporal graph for each patient. Then, From the frequent disease-process subgraphs, the Constrained Frequent Subgraph Mining (CFSM) algorithm is used and then the subgraph is constructed. From the subgraph, Construct the base disease progression network by aggregation of the recoded graph.

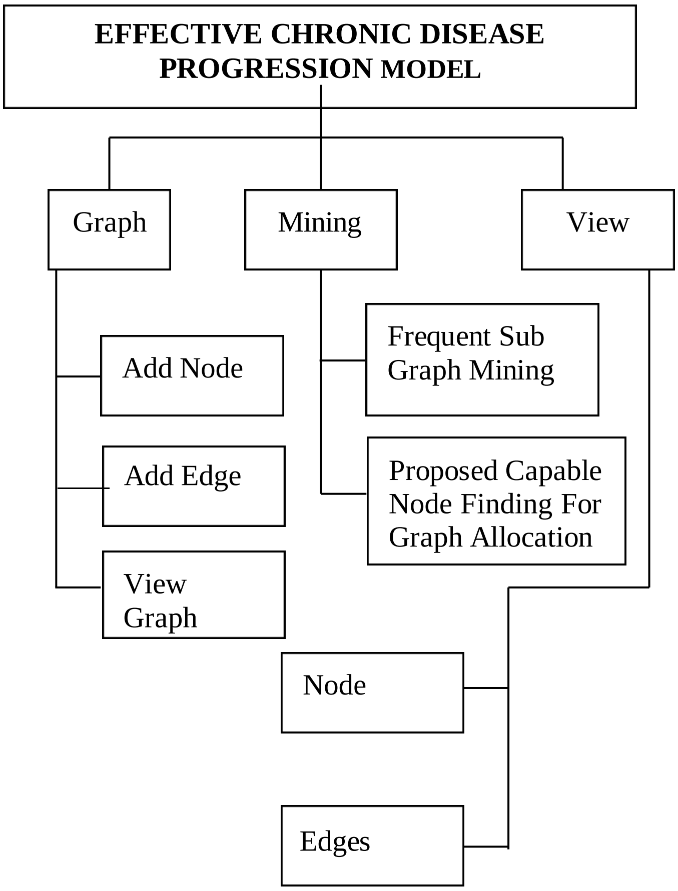
Conduct community detection on the base disease progression network and transform them into chronic disease progression rules. Conduct chronic disease-based healthcare insurance fraud detection according to the rules obtained in Step 4. Additionally, each nodes are assigned with graphs for Map process. All the nodes should be equally balanced by using load balancing technique.

For example, in subgraph one graph is given to Node A and another graph is given to Node B. So, the map process will complete in smaller intervals in all the nodes so that reduce phase can be started immediately.

4.1 ADVANTAGES :

- Before sending input graph data to nodes, they are balanced. Each should nodes should have equal number of nodes and edges.
- Nodes complete the mapper process in smaller intervals. So that Reduce phase can be started quickly without any delay.
- Overall time efficiency is increased.

4.2 SYSTEM FLOW DIAGRAM


EFFECTIVE CHRONIC DISEASE PROGRESSION MODEL
Graph
Mining
View
Add Node
Add Edge
View Graph
Frequent Sub Graph Mining
Proposed Capable Node Finding For Graph Allocation
Node
Edges

5. CONCLUSION :

The proposed system approach aims in calling map functions with more number of key, value pairs and also to get new key, value pairs. And the reduce function is used to compute the new cluster centers.

Unlike existing approach where all the data are given to map functions in single pass, here the passes may or may not continue based on previous iteration values.

6. REFERENCES :

[1] G. Liu,M. Zhang, and F. Yan, “Large-scale social network analysis based on Mapreduce,” in Proc. Int. Conf. Comput. Aspects Soc. Netw., 2010, pp. 487–490.

[2] J. Dean and S. Ghemawat, “Mapreduce: Simplified data processing on large clusters,” Commun. ACM, vol. 51, pp. 107–113, 2008.

[3] U. Kang, C. E. Tsourakakis, and C. Faloutsos, “Pegasus: A petascale graph mining system implementation and observations,” in Proc. 9th IEEE Int. Conf. Data Mining, 2009, pp. 229–238.

[4] U. Kang, B. Meeder, and C. Faloutsos, “Spectral analysis for billion-scale graphs: Discoveries and implementation,” in Proc.15th Pacific-Asia Conf. Adv. Knowl. Discov. Data Mining,2011, pp. 13– 25.