Volume 4, Issue 10, October 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Analyze Different Data Structure for Weighted Frequent Patterns Mining

Ms. Shinde Rupali R, Prof. Maral Vikas B

Dept. Computer Engineering, K.J. College of Engineering and
Management Research, Pune, Maharashtra India

***Abstract:** There are various algorithms for the frequent pattern mining in the current trends but only frequency of item is not so essential for the data mining.Weight is very important for the business analysis.Weight is nothing but the specific value of the item.so we want new techniques for weighted frequent pattern mining.Those algorithms will consider both entities i.e Frequency as well as weight for the data mining.in our paper we have to studied various data structure for arranging the frequent and weighted patterns for mining. Data from data warehouse will regularlychange i.e. updated so we want the algorithms which are adoptable for that updation as well as changing minimum support of the user.But in the their various algorithms are invented but they may be only related to frequent pattern mining and required more than one scan for the mining.In our paper we will study about various data structure for manage weighted frequent pattern for data mining.The data structure will help for the finding the candidate item from the data warehouse instead of association rules,Aproiri based algorithms.which required more than one scan for data mining.*

***Keywords:**Data mining, data structure , various types of information.*

I. INTRDUCTION

The "Knowledge Discovery from data warehouse" process, an interdisciplinary subfield of computer science and Engg., is the computational process of analysis patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall aim of the data mining process is to gather exact related information from a database and rearrange it into a formal structure for next use. Aside from the data analysis step, it involves big dataset and data managementaspects, data preprocessing, model and inference considerations interestingness metrics,complexity considerations, post-processing of discovered structures, visualization, and online updating.

In short data mining is mostly used to mine any form of very big data or information processing (collection, extraction, warehousing, analysis, and statistics) but is also create to any kind data format for computer decision support system, including artificial intelligence, machine learning, and business intelligence. The main data mining task is the automatic recognitionof big quantities of data to gathered previousnot known major patterns such as groups of related data records (cluster recognition), unwanted records (anomaly recognition) and dependencies (association based mining). This usually involves using database techniques such as spatial index. These patterns can then be seen as a kind of over view of the input data, and may be used in next analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might detect various more than one groups in the database , which can then be used to obtain more error free conditional out comes bya decision support system. Not only the data collection, data preparation, but also result computation and reporting are part of the data mining step, but perform the overall KDD process as additional steps.

II. CONCEPT OF DATA MINING AND DATA STRUCTURE

In the above section we have to see the meaning of data mining. Now we are going to see about data structure in the computer science there are various data structure i.e. Array,graph,tree,linkedlist,Hash,Binarytree,Space partition tree etc.Data structure is nothing but the particular way of storing and organizing data in a computer. Data structure provides a mean to manage data veryefficiently.The data structure provides the mapping which may causes time complex1ity and space complex1ity is very less. In the data mining we will discover the data from verylarge data base so the data mining is the application of Business enterprises.As per above definition we have to access data from big data by using various algorithms e.g Map-Reduce algorithms etc. For data analysis we have to use various tool i.e. association rule,machine learning tools,Artificial intelligence etc.But very important point is that how to arrange basic data in data warehouse.So we are going to use various data structure for arranging data in data warehouse. In the Data mining is discovery of data depends on the user request.

III. RELATED CONCEPT

In the data mining we will mine various types of data e.g.Businessdata,organization data, Hospital data, Research data, Spatial data, sequentialdata, webanalysis, sequential dataetc. So we will see various data structure for various application.

3.1 Spatial data mining

In the spatial data mining indexing data structure is used which will speed up the mining process. Geographic Information System is very important for updatations on geographical related information.Geographical Data composed as follows[8]:

a) Spatial Attributes e.g. Coordinates,geometry
b) Non spatial Attribute i.e. name of town

Spatial data in GIS can be represented as Raster and vector model.In the raster model each pixel is act as the index.In the raster model space divide in to grid and each cell act as pixel.In the vector model the basic primitive is points.Data structure is used to store the following concepts of space[1]:

1. point
2. chain
3. polygoan

Fundamental operation for data are

1. Find distance between two object
2. Area of object
3. length of object
4. intersection of object.

3.2 probabilistic Data structure for Web analysis and data mining:

There are 12 probabilistic data structure which are used in the data mining that are bloom filter, Count-minsketch , kinetichanger , heater , Locality sensitive hashing, Min Hashing, Quotientfilter, Random binary tree, randomtree, skiplist, treap.

Now a days data mining is very vital concept of the computer engg. So we want analyze very large amount of data for the mining.But for the powerful data distributed data store like Hadoop and some algorithms are required for the mining.

Probabilistic data structures is described in the following section

- For some structures Bloom filter, in the bloom filter determine parameter of on the basis of calculated value and required error.
- In the Count-Min probability data structure we have to identify the complicated dependency on statistical properties of data and experiments are the only proper way to understand their applicability in day today life.Theapplicability of the probabilistic data structures is not only depend on any specific queries or on a single data set.
- The contrary, structures populated for arranged or process the various types of queries on data sets.

3.3 Frequent Patterns Mining in Data Streams at different Time slot dynamically:

Frequent pattern means support of the pattern .But in stream mining it difficult because any of the infrequent pattern may become frequent in dynamic transaction.

It easy to maintain the data base if the it is static DB. But for the dynamic DB we have to use time-sensitive pattern.

We have to maintain updated tilted-timewindows for each pattern at different timeslots. Now new invention to use the FP-Tree which gives us an great data structure for support pattern mining, [12] we design FP-stream data structure , an effective FP-tree-based model for mining frequent patterns from data streams. An FP-stream structure consists of the following -

(a) A support pattern-tree to recognize the frequent and sub-frequent itemset information.
(b) Atiltedtime window chart for each high supported pattern. Proper algorithms for building, maintaining and updating an FP-stream structure over data streams are ex1plored.

Recently we have some application on the basis of data mining in data stream at different time i.e Network traffic analysis , web click streaming, sensor network analysis power consumption mechanism.In that example continuously tracing the updatation on the data warehouse.Stream data. Stream data management and stream query processor are under the development.Now a days we have to use Landmark model which mine thestream from particular point still point.

3.4 Data mining using index structure in clustering techniques:

In the index structure mining the various related objects and a distance metric and map that object with k-dimensional space , [6]because of that we can preserve proper distance between relevant object. So index structure is useful for the important characteristic of data analysis i.e clustering and visual application . previously there is one algorithm was developed Fastmap but it is related to Euclidean distances clustering ,but metricmap index structure mining is releted to non-Euclidean distances (i.e., general distance metrics).

We have some comparison regarding Fast Map and Metric Map-

i) Fast Map gives a lower relative error than Metric Map for Euclidean distances.

ii) Metric Map gives a lower relative error than Fast Map for non-Euclidean distances (i.e., general distance metrics).

A general distance function isEuclidean distance satisfies following properties.
â that takes a pair of object satisfying the following properties: for any objects x1, y1, z1, ***â*** (x1, x1)=0and ***â*** (x1, y1) > 0,x1 = y1(nonnegative definiteness);***â***(x1, y1)= ***â***(y1,x1) (symmetry); ***â***(x1, y1) ≤ ***â***(x1, z1)+ ***â*** (z1,y1) (triangle inequality).

3.5 Data mining on graph based data structure:

In general a graph (G) is represented as G(V,E) where V is set of node, and E is a set of edge or ink which connect vertices.

Many of the business,artificial intelligence techniques , scientific application the required patterns are very complicated than frequent pattern e.g biometric gen structure , so the developer required extra efforts to mine such type of data.This type of sophisticated data will be managed in various patterns like sets,sequences, tree,network,graph.

In this case graph structure is used because is easy to sophisticated application. With the broad application including bioinformatics, text retrival, web analysis, computer vision etc. Mining frequent subgraph can also work in pattern recognition,classification and clustering. The graph links various nodes together may generate different patternlike telecommunication network,computer network etc.

3.5.1 The graph based mining in the social network:

In general, a graph *G* is represented as *G*(*V*, *E*) where *V*is a set of vertices (or nodes) and *E* is a set of edges (or links) connecting some vertex pairs in *V*.

In this section we have to seen some approaches of online social network.so first of all study on SN (Social Network) and OSN(Online social network).In the social network each user is created a profile for own identity.The social site provide the facility of chat ,sharing, photo album, by using that user can share the messages to the available friends and others of the contact list. But previously both the parties have accept the request of each other and create the link.[2]In the online social network we just introduce two data set which are used in social network that are Data log of computer, data log of online social network.Now we cover key themes that is graph mining. In the graph mining that can be consider a specialization of data mining which is difficult for human being to find meaningful and interpreted knowledge from large data set. The process of extract the data from 1 terabyte DB of transaction to identify the fraudulent pattern we have to use graph theory. Because graph has some properties I,e data representation and interpretation by using various techniques and finely we can discovered the required pattern.In the Fig1. The basic topic are divided as graph,social network dataset, OSN. And Hot topics are communities ,recommendation ,some mathematical model , calculas approach ,behavior , and information diffusion are the input to the graph mining according to that we can generate the required pattern model form large data set.

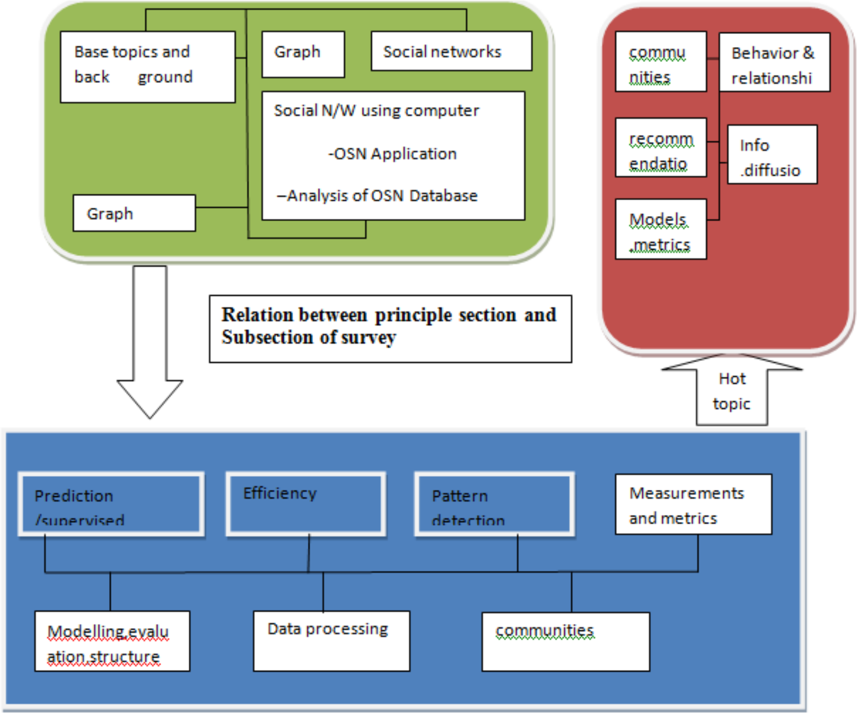

Fig1: Graph based mining in the OSN(Online Social Network.)

IV. REVIEW ON DATA MINING ON TREE BASED DATA STRUCTURE

In the data mining the data are extracted from data warehouse and arranged in the tree like structure. The tree like structure can be used for the data like web analysis,network sensor analysis,sharemarket,various malls , multimedia , data stream ,DNA etc.

4.1 P-tree and association mining for gens expression profiling on DNA:

In the gene expression the researcher are developing the computational tool for the examining the interacting and interpreting result from different organism.In the topic we have some comprehensive approach containing data from microarray experiment performed on different behavior i.e hypoxic and anoxic stress.And data are represented in the form of P-tree.

In that each spot of microarray is represented as pixel with its red and green band.each band is preserve separately in eight recognized form with levels[3].

The genetic constitution of any organism (k) are represented the total number of genes under the two different groups .The X group constitute contains(X1,X2….Xn) and Y group constitute contains (Y1,Y2,..Yn).and they can be represented as 0,1 that is absent and present respectively.That level are arranged in the p-tree.

4.2 Weighted frequent pattern mining:

In the weighted frequent pattern mining the data set are arranged in the tree based on the weight i.e some specification of any data set e.g price, web page click etc,or may be arranged in one specific order of frequency.There are various algorithm are developed for the weighted frequent patter mining.All of these are the association based or aprioribased.In all that cases the itemset are managed in the tree based structure.Cp tree structure is also used for mining. But when the data set are increased or updated there is some problem of re-tree generation.Multiple scanning of data set arerequired.So it needs more time. In the weighted frequent pattern mining the weight dependes on the application or based on the DB.e.g In the in the trading business price is the weight of the product.In day todayy the weight of the item is very important but the weight or confidence are depend on the any property of the data set.

e.g if we consider item or product then we can calculate weight from the price.Another one example that is click on any link in the web page.

Let I(it1,it2,…it_n) are the item set,D be the Transaction DB(t1,t2,………t_n), is the DB is subset of I.Each transaction contains N-item set.

Weight of the pattern is:sum of the weight of all items of pattern/length of p

e.g. p(ab) = weight(p)=weight of (a)+weighted of (b)/length of p

wsupport=weight of pattern*frequency of the pattern.

Table 1:Transaction table

TID	TRANSACTION
T1	a,b
T2	a,c,d,e
T3	c,e
T4	a

In the Table 1 we have mention the transaction i.e suppose we have a,b,c,e items or product. If any of the customer come purchase the item that thing we will say transaction that will indicate T1.In the above transaction table we have to enter all the transaction.

Table 2: Weight table

TID/ITEM	A	B	C	D	E	Tran.Utility	ITEM	PROFIT(PER UNIT)
T_1	2	2	0	0	0	34	A	2
T_2	3	0	12	4	2	88	B	15
T_3	0	0	15	0	3	66	C	3
T_4	4	0	0	0	0	8	D	8
T_5	0	10	0	8	9	277	E	7
T_6	0	7	3	0	4	142		
T_7	1	0	2	0	1	15		
T_8	2	0	0	1	3	33		

In the Table 2 we mention the weight in the form of profit per unit.And transaction utility means total profit of any particular transaction e,g in the transaction table T1 transaction is havint the item a,b.
Transaction utility=(2*2)+(2*15)=34.

V. COMPARATIVE STUDY

- A result of index structure for data mining and clustering

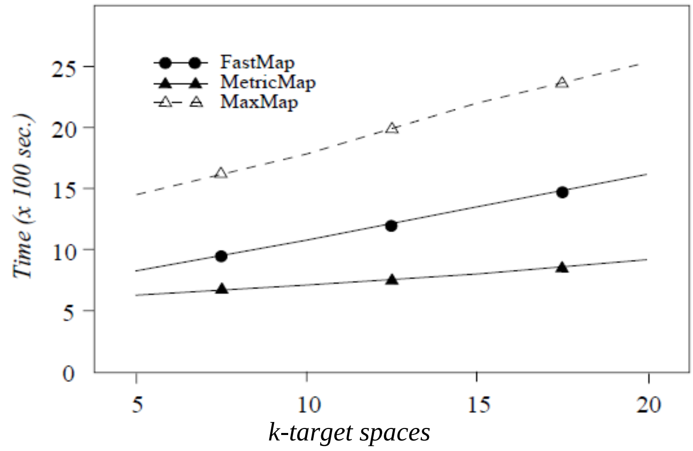

Fig 2: Times of the mappers as a function of the dimensionality of the target space for synthetic Euclidean data.

In the Fig 3 we have to calculate the time in second for finding the associated target spaces for the cluster.According to graph we have to analyze various algorithms in the index structure mining.
Image and various data type data mining using P-tree

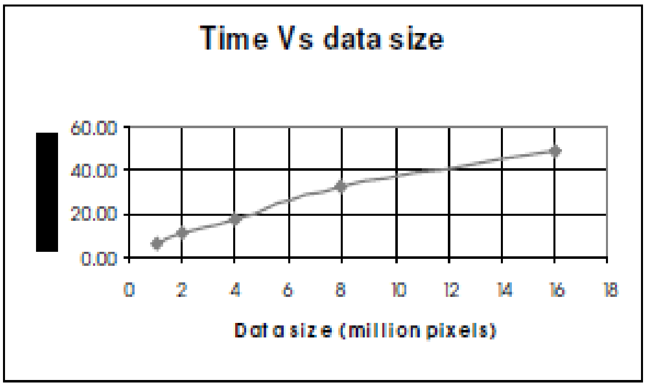

Fig 3:data mining for FP-stream

In the Fig.3 we have to draw the graph on the basis of data stream and time required for the data mining.Data stream is nothing but the continuous data.

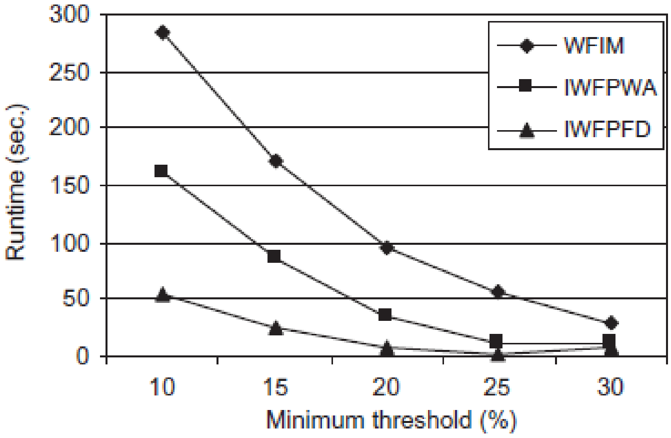

Fig.4 Result on the basis of tree.

In the above Fig.4 we have to draw the graph on the basis of tree data structure for the WFIM(Weighted frequent incremental mining). And in the next algorithms we arrange the weight and frequency in one particular order and draw the graph time Vs Minimum threshold. According to graph WFIM need more time than other two minimum threshold decreases.

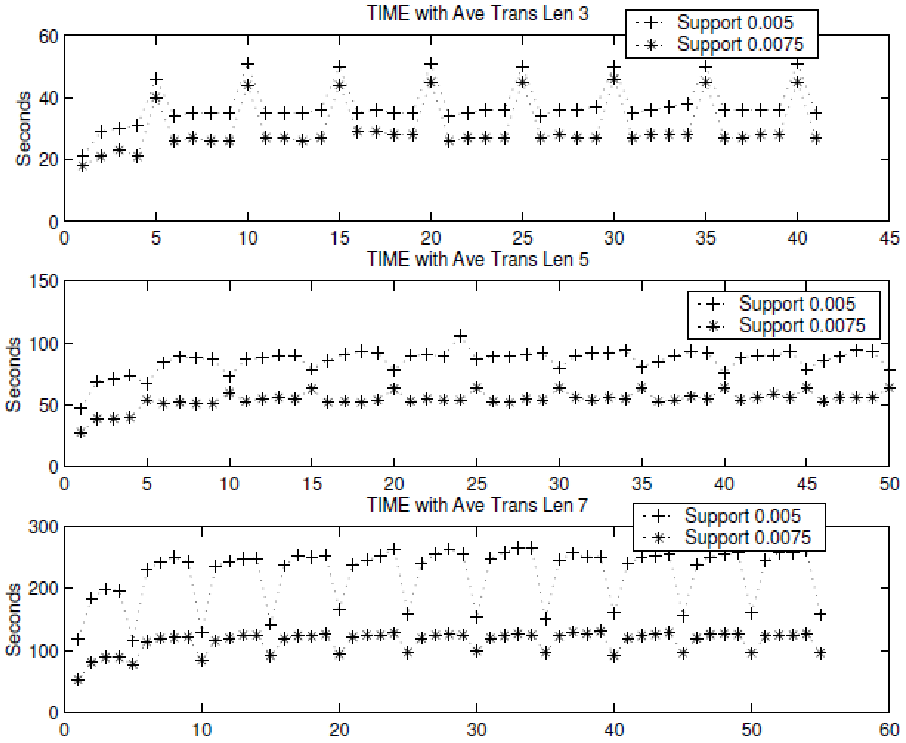

Fig.5 FP-Stream Time required.

In the Fig.5 we have to perform operation on two dataset and that are two supports 0.005 and 0.0075. In the fig5.x-axis shows number batch , y-axis shows times in seconds of per batch. And the transaction length is average 3,5,7. As per the above diagram time and space of the algorithms is changed according to average length and support.So the Mining support Patterns in continuous data at Multiple Time slot is difficult because of average length of the transaction, and support.

VI. CONCLUSION

Till in our paper we have seen various data structure for arranged data set for data mining. Our proposed algorithms is based on the tree based data structure.But in that algorithms we have pass the pattern on the base of weighted as well as frequency within single pass.because the tree structure is very suitable for the incremental and updated data mining. According to above comparisontree structure is best with respect to time when the number of transaction increased. No rescanning is required.And the tree structure is full fill the property of *built once and mine many.* Means in the tree like structure we just want to generate tree at once and use many times.no need to regeneration of tree is required.

REFERENCE

[1] Data mining of social networks represented as graphs and the author is "*UniversitatPompeuFabra, Barcelona, Spain IIIA-CSIC, Bellaterra, Spain"*

[2] data mining on the basis of graph *"Diane J. Cook and Lawrence B. Holder, University of Texas at Arlington"*

[3] ptree& association for gene expression for profile of DNA.

[4] Data Structures for Spatial Data Mining*"PetrKuba"*

[5] data structure"*RuomingJin,gaganagrwal."*

[6] Index based mining and clustering"*Xiong Wang1, Jason T. L. Wang2, King-Ip Lin3Dennis Shasha4, Bruce A. Shapiro5, Kaizhong Zhang6"*

[7] weight frequent pattern mining.*"Unil Yun, B.S."*

[8] Spatial data mining"PetrKuba".

[9] Graph based data mining.*"TakashiWashio,HiroshiMotoda"*

[10] Online Social network on graph based.

[11] Multimedia data mining*"williamperrizo, williamjockheck, amalperera, dongmeiren, weihuawu, yizhang"*

[12] Data mining in data stream.*"ChrisGiannella , Jiawei Han , Jian Pei , Xifeng Yan, Philip S. Yu"*