An Efficient Pruning Technique for Mining Frequent Itemsets in Spatial Databases

G.Parthasarathy[1]
Research Scholar, Dept. of CSE
Sathyabama University, Chennai, India
amburgps@gmail.com

D.C.Tomar[2]
Professor, Dept. of Information Technology
Jerusalem College of Engineering, Chennai, India.
dctomar@gmail.com

***Abstract*— Frequent Itemset Mining is evaluating the rules and relationship within the data items are optimizing it, in the large spatial databases (for e.g. Images, Docs, AVI files etc).It is one of the major problems in DM (Data mining) domain. Finding frequent item set in the large set is one of the computational complexities in mining. To improve the efficiency and performance of the mining frequent item set algorithm, the key term is to apply pruning techniques which reduces the search space and its complexity of the algorithm. Here we proposed a robust technique of pruning called SP pruning for uncertain data's. Here our methodology is used to mine the data sources of uncertain data model. We have analyzed and implemented all well known algorithmic models for mining frequent item sets for both binaries and uncertain data's. Our experimental results show that FPgrowth performance is high for binary data sets where our method performs at high rate of accuracy for uncertain data sets.**

Keywords- *Data mining, SPpruning , FPgrowth, Pruning, Classification*

I. INTRODUCTION

In recent years, data mining has reached in all the domain for storing, processing, retrieving data streams and mining frequent data sets is said to be leaning. At an instance the real world problem is defined implicitly by arrival time and time stamp for explicitly. The algorithm for mining data stream should course in single pass for the characters of data streams. According the survey *FPMAX** is fastest one for all data sets. The algorithm process with two processes, initialization with scanning datasets for constructing FP-Tree.

The main motto is to survive on market business analysis. Consider this example, When in a store where buyers(customers)comes and buys some items, how likely she would buy the specific items or how various buyers buy the items together?.Recently the researchers started to analysis the rules for data where the existence for the particularity is uncertain. For example consider the spatial database SP with various data sets such as multimedia files, docs etc with a black box algorithm BA.BA takes the input .avi files & .png, .jpeg files and deliverables be the appropriate files with the suitable formats(say images or video files)along with their corresponding likelihood values. The output of BA may depend on various key parameters such as frame rate, fps(frames per second),size, format etc. The result may be tabulated and is consider as one of the transaction. Let $I(k) = \{i_1, i_2, i_3, i_4, i_5, i_6 .... i_k\}$ be the set of the item sets called items. A subset X by which $X \subseteq I$ is called as item set. Item set with *K item* is called *K-item* set.

Analysis of these types of data sets may be used to classify the multimedia files to a specific field. For example, a file with high likelihood values for high frame rate, size belongs to the video files and a file with low likelihood values for size, frame rate belongs to image files.

Table 1: Example of multimedia files database.

Resource	Video	Audio	Image
R1	.avi-30%	.mp3-40%	.jpeg-50%
R2	.mpeg-40%	.wmv-45%	.png-45%
R3	.mkv-58%	-	.tiff-55%
R4	.3gp-25%	-	-
R5	.divx-80%	.codec-45%	-
R6	.mp4-56%	-	-
R7	.3gpp-10%	-	-

II. RELATED WORK

This paper, has been developed for digging out and review huge knowledge about the data provided by Deep web. The *deep web* submits to data sources with backend databases that are only accessible through the query forms and Sentiment analysis; where online reviews and blogs can be dig for recognize writers' views and feelings on topics of current interest. The prototypes in differences between the values for the same entity are grouped by inheriting differential rules. In this solution, a statistical hypothesis test had been used to identify important variation in values of the final quantitative attributes between two data sources . The pruning method is used to identify differential rules and pruned if their behaviours are predicted by their complementary ancestor rules. In this algorithm, a hash table is used to store the identified differential rules according to their profile representations and applied four travel related deep web data sources [1]. Since youngxin proposed the very optimal mining algorithm using depth-first search method to acquire entire probabilistic frequent closed itemsets. To shrink the search space and ignore unnecessary computation, probabilistic pruning and bounding methods [4].

The problem of pruning technique using voronoi diagrams to reduce the number of distance calculation is proposed in this paper and improved the effectiveness [5]. All the previous studies on purning technique analysis a data model to reduce the redundant of data.To improve the efficiency of the U-Apriori algorithm; in this paper propose a data trimming technique to ignore irrelevant candidate support increments performed in the Subset-Function. In which, the input and output cost has been reduced [6]. In this paper, we studied the discovery of frequent patterns and association rules from probabilistic data under the Possible World Semantics and which discovers frequent patterns in bottom-up and top-down approaches [7].

III. SPATIAL ARCHITECTURE

In this part we plan a methodology to extract frequent model of spatial objects. These spatial objects situated near to each other for a given sample space of geographic area. The framework can be described as a series of methods. It will dig out the spatial objects and its frequency of availabiligy from the Map Database and builds a model Spatial Objects Datasets. The next method is to first we characterize the frequent order catalog in form of numerical illustration where each object in the transaction is represented Then the next step tree using the numerical illustration of each transaction dataset. The final step finds frequent spatial model with their respective support count by intersecting its numerical ordered list.

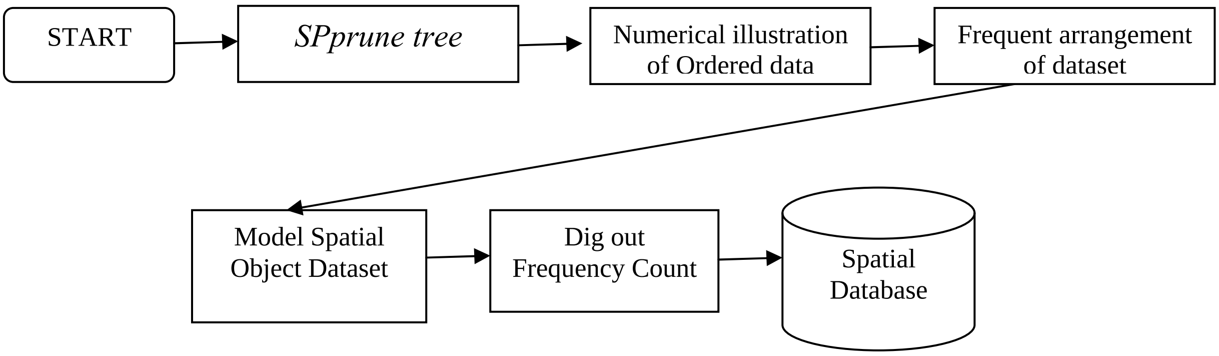

START
SPprune tree
Numerical illustration of Ordered data
Frequent arrangement of dataset
Model Spatial Object Dataset
Dig out Frequency Count
Spatial Database

Fig. 1 Spatial Methodology

IV. TECHNIQUES AND ALGORITHM

Theorems:

Theorem 1: Subset Frequent

Definition: *A subset which is of restricted belongs to any frequent itemset is a maximal frequent item set.*

Theorem 2: Superset Infrequent

Definition: *A subset of any infrequent item set is not a frequent item set.*

Theorem 3: Superset Frequent

Definition: A superset of any frequent item set is a frequent item set.

Theorem 4: Subset Equivalence

Definition: *Let Z & X be the itemsets associated with node n and x. If x is left to the node of n, then* $Z \sqsubset X$

Then $Z \equiv X$ *is defined as Sub (Z) = Sub(X),prune all the children of n.*

Theorem 5: *Subset Infrequent*

> ***Definition:*** *A subset which is of restricted belongs to any frequent item set is a not a maximal frequent item set.*

Pruning Techniques:

Here we use three existing pruning techniques used by *FPMax**[2] namely

- Subset infrequency pruning [1,2].
- Superset frequency pruning [1,2].
- Subset equivalence pruning [1, 2].

A. Subset infrequency pruning:

Consider the node n is in search space tree, then each item *x* in *conn_tail(n)* become the member of *free_tail(n)*,count the support of the item sets *head(n) U {x}.*

```
Begin
        If (Count (head (n) ||x) &&Sup (head (n) ||x))
                Add=free_tail (n);
                head (n)=1
                x=Val (add);
                Assign=head (n)*x;
                Apply theorem2
End
```

Note: Theorem2: A subset of any infrequent item set is not a frequent item set.

B. Superset frequency pruning:

Look ahead Pruning is also called as superset pruning, considering node *x*, if item sets *head(x) U tail(x)* is frequent. Then all the children nodes of x should be pruned according to the theorem 1.There are two methods in existing methodology one is to count the support (breadth first search)and another method is to check for any supersets(depth first search)[2]

C. Subset equivalence pruning (SEP):

SEP is said to be true when it satisfies and obeys the *theorem 4*.

Note: Let Z & X be the item sets associated with node n and x. If x is left to the node of n, then $Z \sqsubset X$ *.Then* $Z \equiv X$ *is defined as Sub (Z) = Sub (X), prune all the children of n.*

Proof:

Let X be any item set associated with the node n, node x be left to the node n, there must be the item $i \in Z \backslash x$ *and* $i \notin X$*.Because sub(Z)*$\equiv$*sub(X),then any transaction Trans which contains X must contain i*[2].

SP pruning:

SPP is said to be true when it satisfies and obeys the *all the theorem 1-5*.

Proof:

Let V be any item set associated with the node n, node x be left to the node n, there must be the item $i \in Z \backslash x$ *and* $i \notin X$*.Because sub(Z)*$\equiv$*sub(X),then any transaction Trans which contains V must contain i*[2].

For example

Fig 4 shows the SPprune algorithm, whereas figure shows the SEP algorithm. Consider the below stated example, set of elements be a, v, c, d for which $\phi \equiv \forall$(theorems)

$v \sqsubset$(a,c),c$\sqsubset$(a,d),d$\sqsubset$(a,v)and

sub(v)=sub(a) U sub(c),sub(c)=sub(a) U sub(d),sub(d)=sub(a) U sub(v)

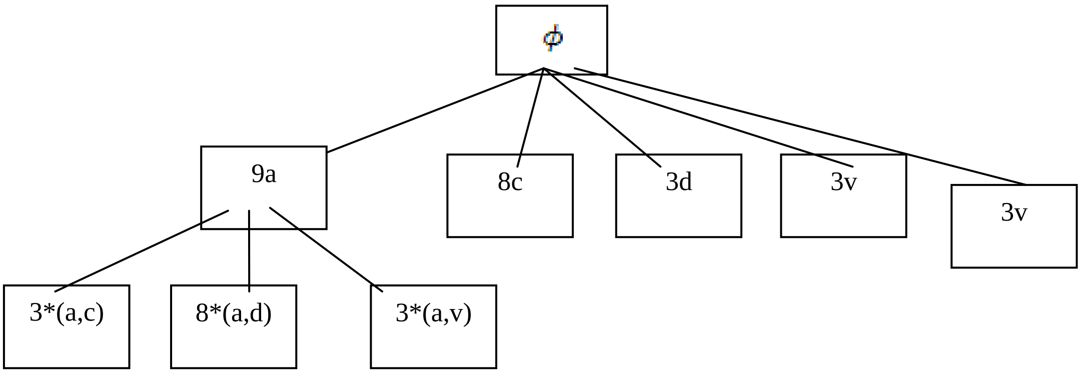


Fig 2: SEP method

Φ

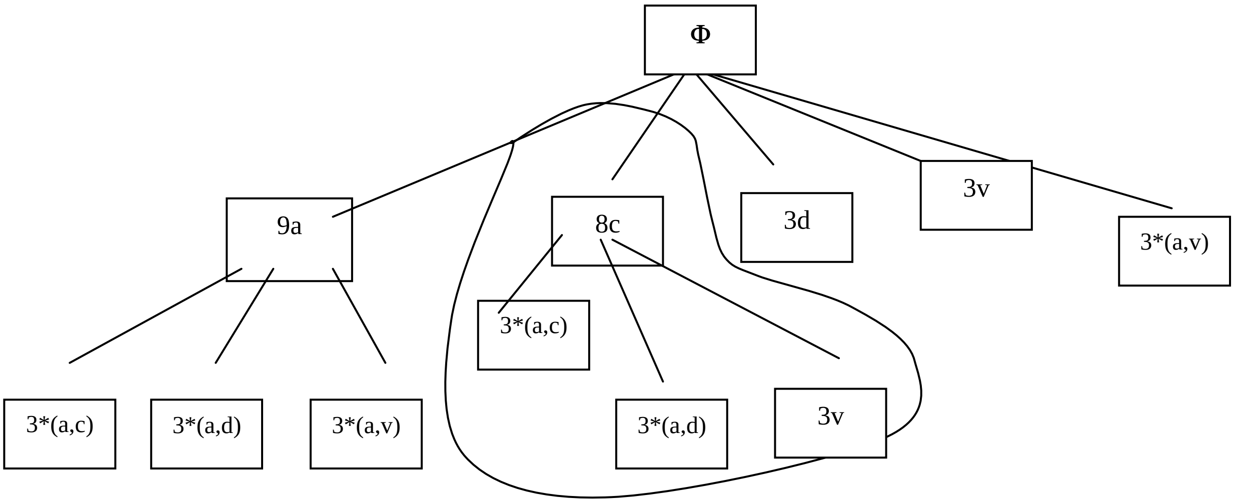


Fig 3: SPPruning method

Mining Frequent Item set:

To support the count of the items, we need to access the records in the database. The database can be stored in the memory on the following ways:

Table 2 : Database Representation

S.No	Representation	Database	Identifier	Scan
1	Horizontal	Transaction Db	Row_id	Single pass
2	Vertical	Transaction DB	Trans_id	Single pass
3	Bit-vector	2D DB's	2d_bitvector	Double pass

FPmax mining algorithm is composed of two main steps namely FPTree construction and mining frequent item sets based on FPTree. FPTree is constructed of mainly two parts: namely 1) Header table & 2) Prefix tree [2].

Header Table:

Header table contains all the items in the DB [2].All link of the header table is set to zero (0).Whenever an item is added to the tree, corresponding entry is updated in the tree.

Prefix Tree:

Prefix tree consists of the entire list of items in Header Table and their support in the prefix tree [2].

Table 3: Transaction items after pre-processing

SID	**ITEMS**	**FREQUENT ITEMS**
S1	A,B,C,D	A,C,B,D
S2	C,D,E	C,D
S3	E,F,A,B	E,F,A,B
S4	C,D,E	C,D
S5	B,E,I	B,E
S6	A,C,E	A,C,E
S7	G,H,E,F	E,F,G
S8	A,G,H,I	H,G,F

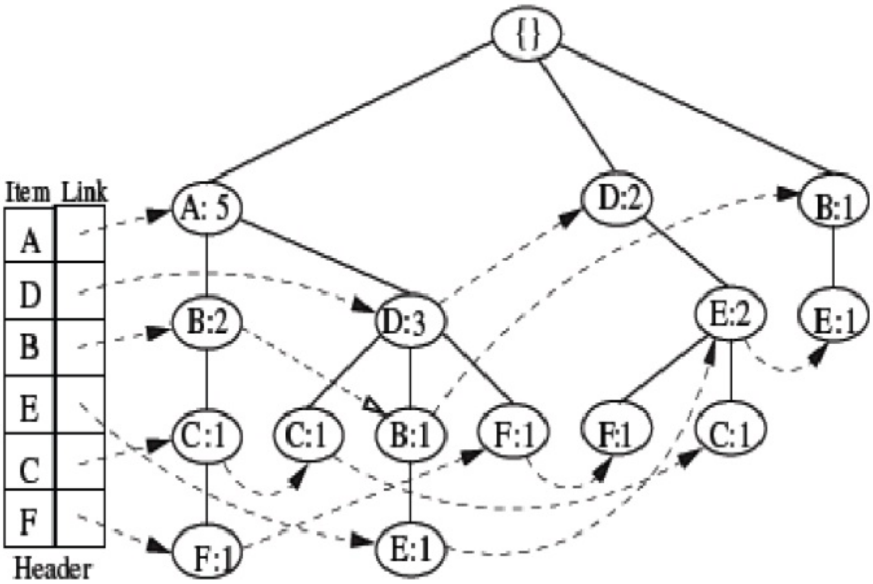


Fig 4: Representation of FPTree

Algorithm:

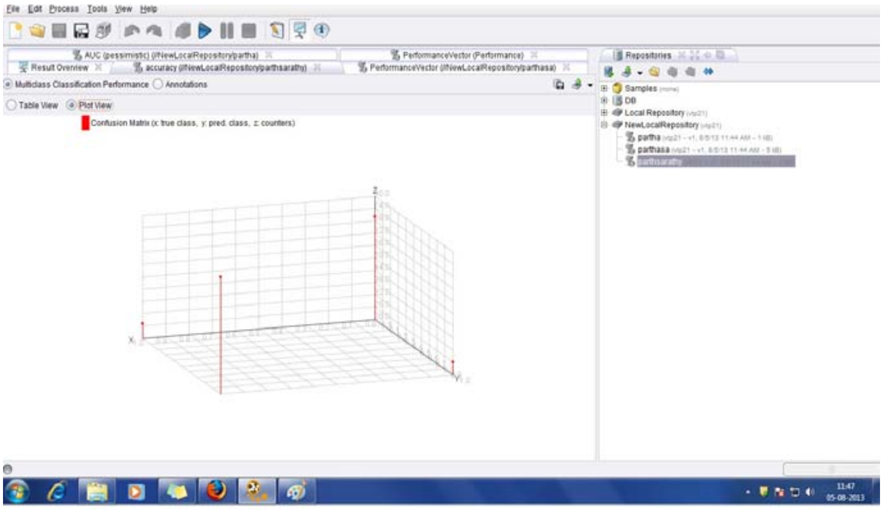
```
Function O= SPprune(T,M,C,Min_sup)
T=input_parameter,M:FPTree,C:MFI for T,Min_sup:CFI for min_sup(SPprune);
Output:O;
If T →→path(p) then
Insert(p)→M;
        Else
        For i:T.header
                If(SPprune≡ ∀(Theorem))
                 Continue;
                        If Sub(i)<min_sup;
                        Set Y=T.base || i;
                        If(sub(i)≡ Y)
                                Add Y to C;
                                If(T[i] !=NULL) then
                                        Free_Tail={Frequent Item set for T[i]}
                                Else
                                Sort_desc(Free_Tail);
                                End if
                        End if
                End if
                Else if (SPprune≡ ∀(Theorem)) then
                        Construct FPTree
                        Goto Insert(FPTree)→M;
                        Merge FPTree with M
                End if
End
    End if
End
```

V. RESULTS & PERFORMANCE MEASURE OF THE ALGORITHM:

We analyze the performance and running time of the algorithm on binary and uncertain data sets. The common dataset taken is *connect* datasets and it was downloaded from [3].The experimental results shows that our proposed methodology can efficient reduce the search space. Here the cost is very low when compared to subset equivalence pruning technique because in our methodology we are focusing our novelty over the frequently used item sets rather focusing on CFI(closed frequent Item sets) and MFI (Maximal frequent Item sets).Here our proposed methodology is to check whether the item sets has superset and their support is equal or not. Here we can store the SP item sets which are lower than the threshold values. Thus the search space of our proposed methodology is at little smaller when compared to other of CFI & MFI. Hence the whole efficiency of the proposed work is at improved rate, especially for uncertain data sets [1, 2, 5, 6, and 9].

Table 4 : Example of Item sets sampling-SPprune

F. ID	ITEMS	FREQUENT ITEMS
F1	A,B,C	A,C,B
F2	C,D,E,F,G	C,D
F3	E	E
F4	A,B,G	A,B,G
F5	B,G	B,G
F6	A,B,C	A,B,C
F7	A,G,B	A,G
F8	A,B,I	B,I

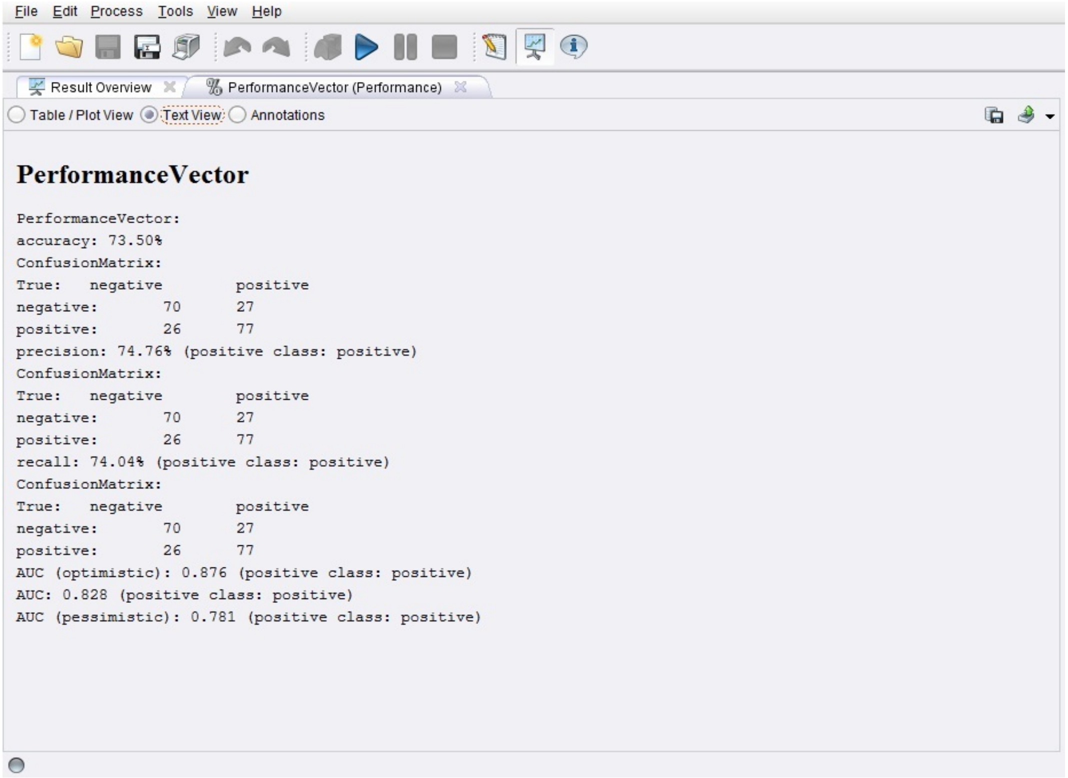

File Edit Process Tools View Help
AUC (pessimistic) (//NewLocalRepository/partha)
PerformanceVector (Performance)
Result Overview
accuracy (//NewLocalRepository/parthsarathy)
PerformanceVector (//NewLocalRepository/parthasa)
Repositories
Samples (none)
DB
Local Repository (vlp21)
NewLocalRepository (vlp21)
partha (vlp21 - v1, 8/5/13 11:44 AM - 1 kB)
parthasa (vlp21 - v1, 8/5/13 11:44 AM - 5 kB)
parthsarathy
x
y
z
11:47
05-08-2013

● Multiclass Classification Performance ○ Annotations

○ Table View ● Plot View

Confusion Matrix (x: true class, y: pred. class, z: counters)

Fig.5.Implementation results of SPprune



PerformanceVector

```
PerformanceVector:
accuracy: 73.50%
ConfusionMatrix:
True:   negative        positive
negative:       70      27
positive:       26      77
precision: 74.76% (positive class: positive)
ConfusionMatrix:
True:   negative        positive
negative:       70      27
positive:       26      77
recall: 74.04% (positive class: positive)
ConfusionMatrix:
True:   negative        positive
negative:       70      27
positive:       26      77
AUC (optimistic): 0.876 (positive class: positive)
AUC: 0.828 (positive class: positive)
AUC (pessimistic): 0.781 (positive class: positive)
```

Fig.6.Performance Vector of SPprune

CONCLUSION

In this paper we extend SP prune for uncertain data. We proved that the proposed method has better accuracy when compared to other methods. Finally we analyzed the performance of frequently used data over binary and uncertain data sets. In binary data sets FPMax* plays better performance than other algorithms while in uncertain data sets SPprune algorithm outperforms than existing other algorithms in most of the cases that we examined. When the support threshold is of 6% -26% we noticed that our algorithm is similar as same as FPMax*.

REFERENCES

[1] Tantan Liu, Fan Wang ,Jiedan Zhu ,Gagan Agrawal ,“Differential Analysis on Deep Web Data Sources”,In 2010 IEEE International Conference on Data Mining Workshops,2010

[2] Fujiang Ao, Yuejin Yan, Jian Huang, Kedi Huang “A Novel Pruning Technique for Mining Maximal Frequent Itemsets”In Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007),2007.

[3] Codes and datasets available at http://fimi.cs.helsinki.fi/.

[4] Yongxin Tong, Lei Chen, Bolin Ding,” Discovering Threshold-based Frequent Closed Itemsets over Probabilistic Data”; International Conference on Data Engineering,2012,IEEE.

[5] Kao B, Lee SD, Cheung DW, Ho WS, Chan KF. Clustering Uncertain Data using Voronoi Diagrams. Data Mining, 2008.ICDM '08. Eighth IEEE InternationalConference 2008, p. 333 – 342.

[6] C. Chui, B. Kao, E. Hung, “Mining frequent itemsets from uncertain data,” in PAKDD, 2007.

[7] L. Sun, R. Cheng, D.W. Cheung, J. Cheng, “Mining uncertain data with probabilistic guarantees,” in KDD, 2010.