

Karnaugh Map Approach for Mining Frequent Termset from Uncertain Textual Data

Abstract—In recent years, uncertain textual data has become ubiquitous because of latest technologies of data collect. As the existing technologies can only collect data in an imprecise way. So in the given paper, we propose frequent pattern mining technique for mining termsets from uncertain textual data. in the said technique, we have conducted a study on uncertain textual data with Karnaugh Map concept. In fact the contribution of given paper is three manifolds. Firstly, we review existing method of finding frequent termset form document data. Secondly a new method UTDKM is proposed for finding frequent termset from uncertain textual data, and lastly, we carried experiments to evaluate the performance of proposed method. The experimental results demonstrate that UTDKM method is better for mining frequent termset in comparison to existing methods.

Index Terms—Uncertain Textual Data, Karnaugh Map, Association Rule Mining, Precise Data.

I. INTRODUCTION

Data mining is the core of knowledge discovery in databases. It is method to find the useful and potential knowledge in database. Knowledge discovery in database (KDD) [1,7,10] lies at the interface of database technology [9], machine learning [10], high performance computing [9], statistics [17] etc. There are many research topics in data mining; the problem of finding frequent patterns[10] from databases has been the subject most of the important topics like association rule mining(ARM). Frequent patterns mining in a sub process of ARM [10,16]. Frequent item set mining is a common data mining task for many real life applications. It helps to generate the previously unknown, potentially useful set of the item which co-occurs. It supports motivated the decision support problems[18] faced by many mining algorithms and focuses for discovering association rules [4].Other algorithm like FP- growth algorithm [10], apriori algorithm [16], Depth-first backtracking [18], Pincer algorithm[17], graph based algorithm [7] etc are also landmark in the area of ARM.// The above said all algorithms extract traditional statics database (like web data, market data etc.), which contains precise data. A transaction (Boolean data mining), or each attribute of a transaction is associated with a quantitative values, however, there are (e.g. areas environmental survey, medical diagnosis, stock market), in which users have uncertain its about the presence or absence of some item or event [2,6] which that needs to be modified in order to handle uncertain data. Fig. 1

shows a tree structure of data mining methods can be classified based on the data.

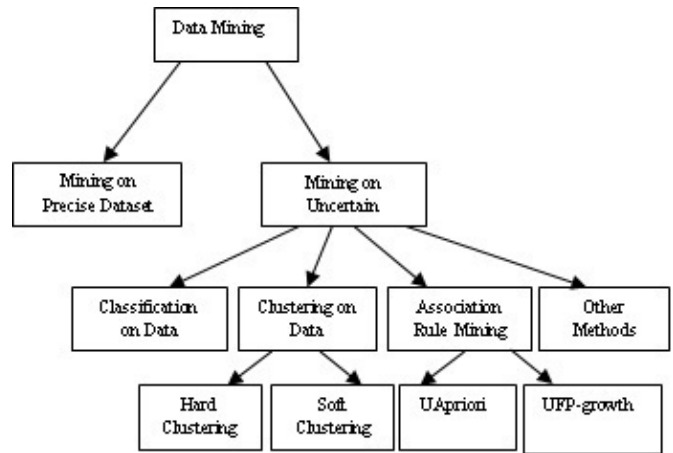


Fig. 1. A Tree Structure of data mining methods can be classified on the data

The concept of a uncertainty measure was introduced by David Appell. A possibility measure on a universe is a function from $(0, 1)$, the basic difference between precise and uncertain textual data is that each document of the later contains terms and their existential probabilities [3, 5, 12, 14]. The uncertainty of such suspicion can be expressed in terms of existential probability. In this data, each term in a document is associated with probability, which indicates the term existences in the document. Normally we can say it is over predication to come the term in the document. Table-I and II shows that the Precise dataset and Uncertain textual Data.

TABLE I
PRECISE DATASET

Document _i d	Terms
D1	A,B,C
D2	B,D
D3	A,B,D

Each Document indicates the chances of that term. the above dataset have one precise dataset and it is also in uncertain form.

TABLE II
UNCERTAIN TEXTUAL DATASET

Document_id	A	B	C	D
D ₁	0.84	0.12	0.04	0.00
D ₂	0.00	0.89	0.00	0.45
D ₃	0.12	0.94	0.00	0.57

Finding frequent patterns from uncertain textual data is not as simple as such like precise data. So, normal approaches that works for precise data are not feasible for uncertain data. There are number of real life situation in which dataset are uncertain, such as sensor data monitoring [24], RFID localization [23] medical dataset[6] and locationbased services [22]hence efficient algorithm for mining uncertain data is in demand [1,3,5].

A. Karnaugh Map

A karnaugh map [13,20] provides a pictorial method of grouping together expressions with common factors and therefore eliminating unwanted variable. karnaugh map reduce the need for extensive calculation by taking advantages of humans pattern recognition capability also permitting the rapid identification and elimination of potential race conditions. A Karnaugh map composed of many grid boxes. Each grid box in a k-map corresponds to a min term or max term. Using the defined min terms, the truth table can be created as a two variables in Table-III and Fig. 2.

TABLE III
TRUTH TABLE FOR TWO VARIABLES

A	B	Results
1	1	T
1	0	T
0	1	T
0	0	F

	0	1
0	F	T
1	T	T

Fig. 2. General Case of a two variables in K-Map

If n is even then matrix of $2^{n/2} \times 2^{n/2}$ is created and if n is odd then matrix $2^{(n-1)/2} \times 2^{(n-1)/2}$ is created.

In this research study we use k-map approach in uncertain textual data to find the frequent termset which reduces the database scans and improve the efficiency and accuracy of algorithm.

II. LITERATURE REVIEW

There are number of different approaches available in literature for frequent pattern mining from uncertain data [1,3,6,12,14,15]. In this section, we have provided some background and discussed related work on data uncertainty. Some researchers have extended association rules mining techniques to imprecise or uncertain data. They have proposed different approaches and framework.

C. K. Chui et.al. was proposed in 2007 Mining frequent itemset from uncertain data. [6]They proposed the U-Apriori algorithm, which is a modified version of the Apriori algorithm, which works on such datasets. They identified the computational problem of U-Apriori and proposed a data trimming framework to address this issue. They proposed framework for mining frequent itemsets from uncertain data. A data trimming framework is proposed to improve mining efficiency. Through extensive experiments, the data trimming technique can achieve significant savings in both CPU cost and I/O cost.

Carson Kai-Sang Leung et. al. was proposed in 2010, Efficient Algorithms for the Mining of Constrained Frequent Patterns from Uncertain Data[3] They proposed U-FPS algorithms to find the frequent patterns for efficient mining that satisfies the user-specified constraints from uncertain data.

C. C. Aggrawal et. al. was proposed in 2009, Frequent Pattern Mining with Uncertain Data[3]. They proposed several classical mining algorithms for deterministic data sets, and evaluated their performance in terms of memory usage and efficiency. The uncertain case has quite different trade-offs from the deterministic case because of the inclusion of probability information.

Peiyi Tang et. al. was proposed in 2011, Mining Probabilistic Frequent Closed Itemsets in Uncertain Databases [14], this paper has defines probabilistic support and probabilistic frequent closed itemsets in uncertain databases for the first time. They proposed a probabilistic frequent closed itemset mining (PFCIM) algorithm to mine from uncertain databases.

Laila A. Abd-Elmegid et. al was proposed in 2010, Vertical Mining of Frequent Patterns from Uncertain Data[12] they extend the state-of-art vertical mining algorithm Eclat for mining frequent patterns from uncertain data generate the proposed UEclat algorithm. In this paper they studied the problem of mining frequent itemsets from existential uncertain data using the Tid set vertical data representation. They also performed comparative study between the proposed algorithm and the well known algorithms.

Wang Kay Ngai et. al was proposed in 2006, Efficient Clustering of Uncertain Data[19] in this paper they studied the problem of uncertain object with the uncertainty regions defined by pdfs. They describe the min-max-dist pruning method and showed that it was fairly effective in pruning expected distance computations. They used four pruning methods, which are independent to each other and can be combined to achieve an even higher pruning effectiveness.

C. C. Aggarwal et.al was proposed in 2008, A Framework for Clustering Uncertain Data Streams[2] they provide a method for clustering. They use general model of the uncertainty in which we assume that only a few statistical measures of the uncertainty are available.

In this paper uncertain textual data is used for generating frequent pattern.

III. PRELIMINARY AND BASIC DEFINITIONS

Definition 1. An termset is frequent if and only if their support value is greater than predefined minimum support ($min\_support$), where $min\_support$ is a user-specified minimum support.

Definition 2. Threshold Probability of an term is $DP(T_1)$ and threshold Probability of another term is $DP(T_2)$ similarity threshold probability of term m is $DP(T_m)$. Then total threshold probability of m item is $DP(T_1, T_2, \dots, T_m) = DP(T_1) * DP(T_2) \dots * DP(T_m) \forall m$ terms that are independent.

Definition 3. Support Counts can be calculated using the following formulas

Calculate l -Termset

$$Supp(T_1) = \{A_{ij} \forall A_{ij} \in (r_3 \cup r_4)/n\} \quad \text{where } i = 3, 4 \text{ and } j = 1, 2, 3, 4 \quad (1)$$

$$Supp(T_2) = \{A_{ij} \forall A_{ij} \in (r_2 \cup r_3)/n\} \quad \text{where } i = 2, 4 \text{ and } j = 1, 2, 3, 4 \quad (2)$$

$$Supp(T_3) = \{A_{ij} \forall A_{ij} \in (c_3 \cup c_4)/n\} \quad \text{where } i = 1, 2, 3, 4 \text{ and } j = 3, 4 \quad (3)$$

$$Supp(T_m) = \{A_{ij} \forall A_{ij} \in (r_i \cup r_{i+1} \dots \cup r_z)/n\} \quad \text{where } T_m \in R_{set} \quad (4)$$

Calculate m -Termset

$$Supp(T_m) = \{A_{ij} \forall A_{ij} \in (c_i \cup c_{i+1} \dots \cup c_k)/n\} \quad \text{where } T_m \in C_{set} \quad (5)$$

$$Supp(T_m T_1) = \{A_{ij} \forall A_{ij} (containerrowsof T_m * containercolumnsof T_1)\} / n \quad (6)$$

Following Table-IV shows the number of possible termsets in the order of terms.

Lemma 1: If two events A and B are independent, since whatever happens to the first event A cannot affect the second event B, so the probabilities are therefore multiplied.

Proof: Let the probability of event A is $P(A)$ and the probability of event B is $P(B)$, then the probability of A and B is

$$P(A \cup B) = P(A) * P(B) \quad (7)$$

TABLE IV
ALL POSSIBLE FREQUENT TERMSETS

Number of terms in the termset	Number of possible combinations	Possible number of termsets
1	${}^m C_1 = 1$	$T_1, T_2, T_3, \dots, T_m$ $T_1, T_2, T_3, \dots, T_m$ $T_1, T_2, T_3, \dots, T_m$ $\vdots$ $T_1, T_2, T_3, \dots, T_m$
2	${}^m C_2 = m!/m-2! * 2!$	$T_1, T_2, T_3, \dots, T_m$ $T_1, T_2, T_3, \dots, T_m$ $T_1, T_2, T_3, \dots, T_m$ $\vdots$ $T_1, T_2, \dots, T_{m-1}, T_m$
3	${}^m C_3 = m!/m-3! * 3!$	$T_1, T_2, T_3, \dots, T_m$ $T_1, T_2, T_3, \dots, T_m$ $T_1, T_2, T_3, \dots, T_m$ $\vdots$ $T_1, T_2, \dots, T_{m-1}, T_m$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
m	${}^m C_m = 1$	$T_1, T_2, \dots, T_{m-1}, T_m$

IV. PROPOSED METHOD

In this paper, we proposed k-map approach that mines from uncertain textual data and find frequent termset, it satisfies the user-specified succinct constraint.

A. Proposed Algorithm

Algorithm
START
Input: $T = \{T_1, T_2, \dots, T_m\}$ // set of all termset $D = \{D_1, D_2, \dots, D_n\}$ // set of all Documents $DP(T)$ is threshold probability of Term T $Min\_support=1$
Output: Set of frequent termsets
Process:
1) Scan document dataset D and compare with threshold probability $DP(T)$ and Create k-map for l -termset.
2) for($x=2$ to m)
3) Calculate all combination of i -termset of all term T and calculate probability for all.
4) Compare the probability of all termset with $DP(T)$.
5) If probability of termset is greater than or equal to the threshold probability then increase the k-map map block.
6) Go to step 2.
7) Calculate the support count for m -termset.
8) Now Compare the support count of termset if $Supp \geq min\_support$ then termset is frequent.

END

B. An illustrative example

Let us consider the following database where $m=4$. The terms are $T = \{T_1, T_2, T_3, T_4\}$. D is the set of document $D = \{D_1, D_2, D_3, \dots, D_{20}\}$.

TABLE V
UNCERTAIN TEXTUAL DATASET D

Document id/ Termset	T ₁	T ₂	T ₃	T ₄
D ₁	0.84	0.62	0.31	0.34
D ₂	0.52	0.81	0.08	0.62
D ₃	0.95	0.73	0.56	0.43
D ₄	0.63	0.98	0.23	0.55
D ₅	0.72	0.57	0.76	0.81
D ₆	0.41	0.91	0.81	0.86
D ₇	0.79	0.84	0.94	0.94
D ₈	0.84	0.95	0.76	0.84
D ₉	0.15	0.76	0.86	0.75
D ₁₀	0.48	0.81	0.79	0.78
D ₁₁	0.95	0.73	0.08	0.62
D ₁₂	0.63	0.98	0.56	0.55
D ₁₃	0.72	0.95	0.76	0.59
D ₁₄	0.84	0.76	0.81	0.84
D ₁₅	0.79	0.80	0.94	0.95
D ₁₆	0.87	0.73	0.81	0.12
D ₁₇	0.76	0.72	0.69	0.91
D ₁₈	0.85	0.97	0.99	0.04
D ₁₉	0.01	0.21	0.64	0.98
D ₂₀	0.11	0.29	0.91	0.86

In the above uncertain textual data which is shown in Table-V, each document contains terms and their corresponding existence probability. For example there are four terms T_1, T_2, T_3 , and T_4 . In the first Document D_1 , where existence probability of their terms are 0.84, 0.62, 0.31, and 0.34 respectively.

In first step, our algorithm Find all combination of 2-termset and there probability set and in same way extended for all combination of 3-termset and find there probability set.

Now find all combination of 4-termset and find there probability set.

Then we calculate all possibility of termset and then expected threshold based on Lemma 1 as given below.

- 1) Compare the probability of all 1- termset with expected threshold value. $\{T_1, T_2, T_3, \dots, T_m\} \geq \{0.80\}$
- 2) Compare the probability of all 2- termset with expected threshold value. $\{\{T_1, T_2\}, \{T_1, T_3\}, \{T_3, T_4\} \dots \{T_1, T_m\}\} \geq 0.64 \{0.80 * 0.80\}$
- 3) Compare the probability of all 3- termset with expected threshold value. $\{\{T_1, T_2, T_3\}, \{T_1, T_2, T_4\}, \{T_2, T_3, T_4\} \dots \{T_2, T_3, T_m\}\} \geq 0.51 \{0.80 * 0.80 * 0.80\}$

TABLE VI
PROBABILITY OF ALL 2-TERMSET COMBINATION

S.No.	Termset	Probability
1	$\{T_1, T_2\}$	{0.5208, 0.4212, 0.6935, 0.6174, 0.4104, 0.3731, 0.6636, 0.798, 0.114, 0.3888, 0.6935, 0.6174, 0.684, 0.6384, 0.632, 0.6351, 0.5472, 0.8245, 0.0021, 0.0319}
2	$\{T_1, T_3\}$	{0.2604, 0.0416, 0.532, 0.1449, 0.5472, 0.3321, 0.7426, 0.6384, 0.129, 0.3792, 0.076, 0.3528, 0.5472, 0.6804, 0.7426, 0.7047, 0.5244, 0.8415, 0.0064, 0.1001}
3	$\{T_1, T_4\}$	{0.2856, 0.3224, 0.4085, 0.3465, 0.5832, 0.3526, 0.7426, 0.7056, 0.1125, 0.3744, 0.589, 0.3465, 0.4258, 0.7056, 0.7505, 0.1044, 0.6916, 0.034, 0.0098, 0.0946}
4	$\{T_2, T_3\}$	{0.1922, 0.0648, 0.4088, 0.2254, 0.4332, 0.3526, 0.7426, 0.7056, 0.1125, 0.3744, 0.589, 0.3465, 0.4248, 0.7056, 0.7505, 0.5913, 0.4968, 0.9603, 0.1344, 0.2639}
5	$\{T_2, T_4\}$	{0.2108, 0.5022, 0.3139, 0.539, 0.3495, 0.5832, 0.7826, 0.7896, 0.798, 0.57, 0.6318, 0.4526, 0.539, 0.5605, 0.6384, 0.0876, 0.6552, 0.0388, 0.2098, 0.2494}
6	$\{T_3, T_4\}$	{0.1054, 0.0496, 0.2408, 0.1265, 0.6156, 0.6966, 0.8836, 0.6384, 0.645, 0.6162, 0.0496, 0.308, 0.4484, 0.6804, 0.893, 0.0972, 0.6279, 0.0396, 0.6272, 0.7826}

TABLE VII
PROBABILITY OF ALL 3-TERMSET COMBINATION

S.No.	Termset	Probability
1	$\{T_1, T_2, T_3\}$	{0.1614, 0.034, 0.389, 0.142, 0.312, 0.302, 0.622, 0.606, 0.09, 0.307, 0.055, 0.346, 0.549, 0.517, 0.594, 0.5144, 0.3776, 0.8163, 0.0014, 0.029}
2	$\{T_1, T_2, T_4\}$	{0.177, 0.2611, 0.2982, 0.3396, 0.3324, 0.3208, 0.6238, 0.671, 0.086, 0.3032, 0.430, 0.3395, 0.4036, 0.5363, 0.6004, 0.076, 0.498, 0.032, 0.0021, 0.027}
3	$\{T_1, T_3, T_4\}$	{0.0898, 0.0258, 0.229, 0.082, 0.4432, 0.2856, 0.698, 0.536, 0.097, 0.296, 0.0471, 0.1940, 0.323, 0.5715, 0.7055, 0.086, 0.4772, 0.3366, 0.063, 0.0860}
4	$\{T_2, T_3, T_4\}$	{0.0653, 0.0402, 0.1758, 0.124, 0.3508, 0.6339, 0.7422, 0.6065, 0.4902, 0.4991, 0.036, 0.3018, 0.426, 0.517, 0.7144, 0.071, 0.4520, 0.0384, 0.132, 0.227}

- 4) Compare the probability of all 4- termset with expected threshold value. $\{\{T_1, T_2, T_3, T_4\} \dots \{T_1, T_2, T_3, T_m\}\} \geq 0.41 \{0.80 * 0.80 * 0.80 * 0.80\}$
- 5) Compare the probability of all m- termset with expected threshold value. $\{\{T_1, T_2, T_3, T_4, \dots\} \dots \{T_1, T_2, T_3, \dots, T_m\}\} \geq DP_m \{0.80 * 0.80 * 0.80 * 0.80 * \dots * 0.80\}$

Let $T = (T_1, T_2, T_3, T_4)$ be the set of all terms, where each term is a probabilistic value . if its value is greater then expected threshold then it can hold 1 either . $D = (D_1, D_2, D_3, \dots, D_n)$ the set of all Documents. Each documents D_n contains a subset of terms chosen From T . In

TABLE VIII
PROBABILITY OF ALL 4-TERMSET COMBINATION

S.No.	Termset	Probability
1	$\{T_1, T_2, T_3, T_4\}$	$\{0.0549, 0.02108, 0.1673, 0.0781, 0.253, 0.260, 0.585, 0.509, 0.0675, 0.2394, 0.0341, 0.1903, 0.324, 0.434, 0.565, 0.0617, 0.3435, 0.0327, 0.0013, 0.0249\}$

association a collection of zero or more items is termed as termset. If a termset contains m terms, it is called m -termset. A k-map table is created with first two bits representing terms T_1, T_2 in the Rows and next two bits representing items T_3, T_4 in the columns, which is shown in Fig.4. Then for each document in the database we can read the terms and can mark 1 in the corresponding row and column of k-map. Next time if same bits are appearing then its value can be incremented by one otherwise place 1 in the corresponding row and column as shown in table. Values present in the k-map shows the frequencies of the items.

T_1, T_2	T_3, T_4			
	00	01	11	10
00	0	9	7	8
01	10	5	5	5
11	8	4	4	7
10	7	5	4	6

Fig. 3. K-Map of term frequency

First we calculate the support count of 1 -termset. They all are the elements of either R_{set} or C_{set} . now we use equation (1) to calculate support value of term T_1 . Then use equation (2) to calculate support value of T_2 term, Let the $min\_support = 1$.

$$Supp(T_1) \rightarrow [(8 + 4 + 4 + 7) + (7 + 5 + 4 + 6)] / 20 = 45/20 = 2.25$$

$$Supp(T_2) \rightarrow [(10 + 5 + 5 + 5) + (8 + 4 + 7 + 4)] / 20 = 48/20 = 2.4$$

$$Supp(T_3) \rightarrow [(8 + 5 + 6 + 7) + (7 + 5 + 4 + 4)] / 20 = 46/20 = 2.3$$

$$Supp(T_4) \rightarrow [(9 + 5 + 5 + 4) + (7 + 5 + 4 + 4)] / 20 = 43/20 = 2.15$$

Now compare the support count of 1 -termsets with $min\_support$ to generate single item. Here all 1 -termsets have $Supp \geq min\_support$. So all 4 terms

$\{T_1\}, \{T_2\}, \{T_3\}, \{T_4\}$ are frequent. Now calculate the support count for all 2 -termsets.

$$Supp(T_1, T_2) \rightarrow [(8 + 4 + 7 + 4)] / 20 = 23/20 = 1.15$$

$$Supp(T_1, T_3) \rightarrow [(6 + 4 + 7 + 4)] / 20 = 21/20 = 1.05$$

$$Supp(T_1, T_4) \rightarrow [(5 + 4 + 4 + 4)] / 20 = 17/20 = 0.85$$

$$Supp(T_2, T_3) \rightarrow [(5 + 5 + 7 + 4)] / 20 = 21/20 = 1.05$$

$$Supp(T_2, T_4) \rightarrow [(5 + 4 + 5 + 4)] / 20 = 18/20 = 0.9$$

$$Supp(T_3, T_4) \rightarrow [(7 + 5 + 4 + 4)] / 20 = 20/20 = 1$$

Generate of 2 -frequent termsets is shown in the $\{T_1, T_2\}, \{T_1, T_3\}, \{T_2, T_3\}, \{T_3, T_4\}$ are frequent termsets and $\{T_1, T_4\}, \{T_2, T_4\}$ are not frequent. because their support value is less than define minimum support. Now calculate the support count for all 3 -termsets.

$$Supp(T_1, T_2, T_3) \rightarrow [(7 + 4)] / 20 = 11/20 = 0.55$$

$$Supp(T_1, T_2, T_4) \rightarrow [(4 + 4)] / 20 = 8/20 = 0.40$$

$$Supp(T_1, T_3, T_4) \rightarrow [(4 + 4)] / 20 = 8/20 = 0.40$$

$$Supp(T_2, T_3, T_4) \rightarrow [(5 + 4)] / 20 = 9/20 = 0.45$$

Now we compare the support counts with $min\_support$ to generate 3 -frequent termsets as none of these termsets has $Supp \geq min\_support$.

Finally to summarize by applying our proposed approach that capture the content of uncertain data in example, we found frequent patterns $\{T_1, T_2\}, \{T_1, T_3\}, \{T_2, T_3\}, \{T_3, T_4\}$ in second level. According to need calculate confidence of each possible association rules.

V. EXPERIMENTAL EVALUATION

In this section, we have presented the performance of proposed approaches with the existing classical frequent pattern mining algorithms of UApriori [1], UH-mine[6], and UFP-growth[3], respectively. We compared our algorithm and their performance with the state-of-the-art frequent itemset mining algorithm for uncertain textual data sets. The experiments were

performed on an Intel core 2 Duo, 2.94 GHz system running Windows 7 professional with 2 GB of RAM and TURBO C++. Fig. 4 shows the performance of our proposed approach with using existing approaches.

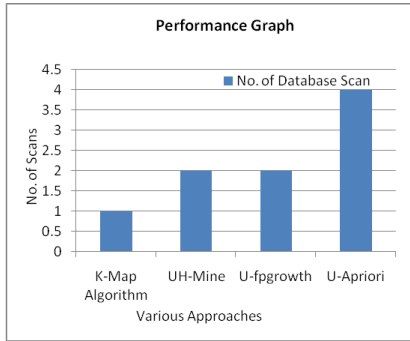


Fig. 4. Performance Comparison with Traditional approaches

Datasets: To test the proposed approach, two different kinds of most popular dataset is used: 20 News groups[8], Reuters[8], which are widely adopted and has become a popular dataset for experiment in document applications of machine learning techniques, such as document classification and document clustering task. Moreover, Table IX summarizes the statistics of these datasets. The detailed information of these datasets is described as follows:

20 Newsgroups: The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups.

Reuters: Reuters is a text document dataset, derived from Reuters-21578 text categorization test collection Distribution 1.0.

TABLE IX
DATASET DESCRIPTION

Dataset	Document	Class
20 newsgroup dataset	20000	20
ReutersText Categorization Collection	8654	52

VI. CONCLUSION

Several data mining methods have been proposed in the last decade. In this paper we studied the problem of finding frequent termsets from uncertain textual data. The proposed Karnaugh map technique performs well for mining patterns (termset) from uncertain data. We introduced a new approach UTDKM for frequent termset, and also compare performance of this algorithm with existing algorithms like UApriori, UH-mine, and UFP-growth, respectively. It needs only one database scan and stores the data in the form of frequency in karnaugh map. The proposed work also shows the importance of high performance gain in terms of both computational cost and I/O cost. The limitation of Karnaugh map technique is that it is very efficient for small number of variables but its performances depreciate exponentially with the increase in

number of variables. Further this work can be extended by the use of QuineMcCluskey algorithm and solve the problem of high dimensional data and the frequent termset may used for document clustering.

VII. ACKNOWLEDGEMENT

This work is supported by research grant from MANIT, Bhopal, India under Grants in Aid Scheme 2010-11, No. Dean(RC)/2010/63 dated 31/08/2010.

REFERENCES

- [1] C. C. Aggarwal *An Introduction to uncertain data algorithm and applications* Advances in Database Systems Volume 35, 2009, pp 1–8.
- [2] C. C. Aggarwal, Philip S. Yu, *A Framework for Clustering Uncertain Data Streams* Data Engineering, IEEE 24th International Conference on ICDE 2008 7-12 April 2008, pp. 150-159.
- [3] C.C. Aggarwal, Yan Li, Jianyong Wang, Jing Wang *Frequent pattern mining with uncertain data* In Proc. KDD 2009, pp. 29-37.
- [4] C.K.-S. Leung, B. Hao, *Mining of frequent itemsets from streams of uncertain data* in Proc. IEEE ICDE 2009, pp. 1663-1670.
- [5] C.K.-S. Leung, C. L. Carmichael, Boyu Hao *Efficient mining of frequent patterns from uncertain data*, In Proc. IEEE ICDM Workshops 2007, pp. 489-494.
- [6] C. K. Chui, Ben Kao, Edward Hung, *Mining Frequent Itemsets from Uncertain Data* Springer-Verlag Berlin Heidelberg ,PAKDD 2007, LNAI 4426, pp. 47-58, 2007.
- [7] D.S Rajput, R.S. Thakur, G.S. Thakur *Rule Generation from Textual Data by using Graph Based Approach* in International Journal of Computer Application (IJCA), Vol. 31 No.9, October 2011 pp 36–43.
- [8] D.S Rajput, R.S. Thakur, G.S. Thakur, *An Integrated Approach and framework for Document Clustering using Graph based Association Rule Mining* in Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28–30, 2012.
- [9] D.S Rajput, R.S. Thakur, G.S. Thakur, *Fuzzy Association Rule Mining based Frequent Pattern Extraction from Uncertain Data* presented in IEEE 2nd World Congress on Information and Communication Technologies (WICT- 2012) October 30-November 02, 2012 pp 709–714.
- [10] Han I and Kamber M, *Data Mining concepts and Techniques* M. K. Publishers, pp.335–389 (2000).
- [11] J. Huang, Lyublena Antova, Christoph Koch, Dan Olteanu, *MayBMS: A probabilistic database management system* in Proc. ACM SIGMOD 2009, pp. 1071–1074.
- [12] Laila A. Abd-Elmegid, Mohamed E. El-Sharkawi, Laila M. El-Fangary Yehia K. Helmy, *Vertical Mining of Frequent Patterns from Uncertain Data* Computer and Information Science Vol. 3, No. 2; May 2010, pp 171–179.
- [13] Neelu Khare, Neeru Adlakha and K. R. Pardasani, *Karnaugh Map Model for Mining Association Rules in Large Databases*, International Journal of Computer and Network Security, Vol. 1(2), 2009, pp.16–21.
- [14] Peiyi Tang, Erich A. Peterson, *Mining Probabilistic Frequent Closed Itemsets in Uncertain Databases* 49th ACM Southeast Conference, March 24–26, 2011.
- [15] Q. Zhang, Feifei Li, Ke Yi *Finding frequent items in probabilistic data*, In Proc. ACM SIGMOD 2008, pp. 819–832.
- [16] R. Agrawal, R. Srikant, *Fast algorithms for mining association rules* In Proc. VLDB 1994, pp. 487–499.
- [17] R.S. Thakur, R.C. Jain, K.R. Pardasani, *Graph Theoretic Based Algorithm for Mining Frequent Patterns* In IEEE World Congress on Computational Intelligence, Hong Kong, (2008) pp. 629–633.
- [18] Rajput D.S, Thakur R.S., Thakur G.S., *Fuzzy Association Rule Mining based Knowledge Extraction in Large Textual Dataset* in International Conference on Computer Engineering Mathematical Sciences (ICCEMS 2012) August 27-28, 2012 pp. 191–194.
- [19] Wang Kay Ngai, Ben Kao, Chun Kit Chui, *Efficient Clustering of Uncertain Data*, Proceedings of the Sixth International Conference on Data Mining (ICDM'06), 2006 pp 2701–2709.

- [20] Yi-Chun Lin, Chun-Min Hung, Yueh-Min Huang, *Mining Ensemble Association Rules by Karnaugh Map* World Congress on Computer Science and Information Engineering 2009, pp 320–324.
- [21] David Appell, "The New Uncertainty Principle", Scientific American, January 2001.
- [22] N. Pelekis, I. Kopanakis, E. E. Kotsifakos, E. Frentzos and Y.Theodoridis. Clustering Uncertain Trajectories. In Knowledge and Information Systems, 2010.
- [23] H. Chen, W. S. Ku, H. Wang and M. T. Sun. Leveraging Spatio-Temporal Redundancy for RFID Data Cleansing. In SIGMOD, 2010.
- [24] A. Deshpande, C. Guestrin, S. R. Madden, J. M. Hellerstein and W. Hong. Model-Driven Data Acquisition in Sensor Networks. In VLDB, 2004.