British Journal of Mathematics & Computer Science
4(3): 333-346, 2014

SCIENCEDOMAIN *international*
www.sciencedomain.org

Karnaugh Map Approach for Mining Frequent Termset from Uncertain Textual Data

D. S. Rajput, R. S. Thakur and G. S. Thakur

Department of Computer Applications, MANIT Bhopal, India

Original Research Article

Received: 18 July 2013
Accepted: 23 September 2013
Published: 04 November 2013

Abstract

In recent years, uncertain textual data has become ubiquitous because of the latest technology used for data collection. As the existing technology can only collect data in an imprecise way. Furthermore, various technologies such as privacy-preserving data mining create data which is inherently uncertain in nature. So this paper propose a frequent pattern mining technique for mining termsets from uncertain textual data. This technique has conducted a study on uncertain textual data using the Karnaugh Map concept. The paper describes the approach in a three step procedure. First, we review existing methods of finding frequent termsets from document data. Second, a new method UTDKM (Uncertain Textual Data Mining using Karnaugh Map) is proposed for finding frequent termset from uncertain textual data. Finally, we carried out experiments to evaluate the performance of the proposed method. The experimental results demonstrate that the prominent feature of this method that is it requires only a single database scan for mining frequent patterns. It reduces the I/O time as well as CPU time.

Keywords: Uncertain Textual Data; Karnauph Map; Association Rule Mining; Precise Data
2010 Mathematics Subject Classification: 53C25; 83C05; 57N16

1 Introduction

Data mining is a method used to find the useful and potential knowledge in a database. Knowledge discovery in a database (KDD) [1], [2], [3] lies at the interface of database technology [4], machine learning [3], high performance computing [4] and statistics [5]. There are many research topics in data mining; one of the most important topic is Association Rule Mining (ARM), which is used to find association between frequent patterns [3] from databases. Frequent pattern mining in a sub task of ARM [3], [6] and has been applied for mining data in many real life applications. It helps to generate the previously unknown, potentially useful set of items which co-occur. It supports the decision

**Corresponding author: E-mail: dharm_raj85@ yahoo.co.in*

support problems [7] faced by many mining algorithms and focuses on discovering association rules [8]. Other algorithm like the FP- growth algorithm [3], apriori algorithm [6], depth-first backtracking [7], pincer algorithm [5], graph-based algorithm [2], [5] are also landmarks in the area of ARM. The above algorithms extract traditional statics from a database (like web data, market data), that contains precise data. A transaction (Boolean data mining) or each attribute of a transaction is associated with quantitative values. However, there are situations (e.g., areas environmental survey, medical diagnosis, stock market) in which users are uncertain about the presence or absence of any item or event [9], [10] that needs to be modified in order to handle uncertain data. Figure 1 shows a tree structure of data mining methods can be classified based on the data.

The concept of an uncertainty measure was introduced by Appell D.. 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 latter contains terms and their existential probabilities [11], [12], [13], [14]. In uncertain textual data, each term in a document is associated with a probability, which indicates the probable termś existence in the document. Table 1 and Table 2 shows the precise dataset and uncertain textual data.

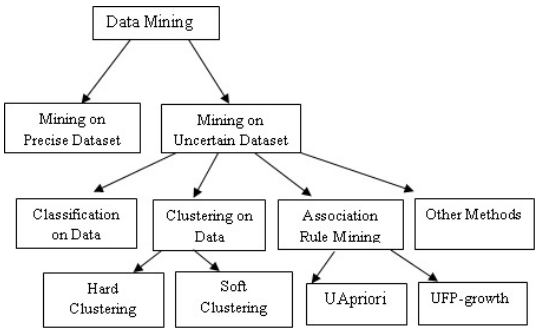


Figure 1: Classification of data mining methods

Each document indicates the chances of that term. The above dataset has one precise dataset and which is also in uncertain form.
Finding frequent patterns from uncertain textual data is not as simple for precise data. So, normal approaches that work for precise data are not applicable for uncertain data. There are a number of real life situations in which datasets are uncertain, such as sensor data monitoring [15], RFID localization [16], medical datasets [10] and location based services [17]. Hence an efficient algorithm for mining uncertain data is in demand [1], [11], [12].

Table 1: Example of precise dataset

Document_id	Terms
D_1	A,B,C
D_2	B,D
D_3	A,B,D

Table 2: 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

1.1 Karnaugh Map

A Karnaugh map [18], [19] provides a pictorial method of grouping together expressions sharing common factors thus eliminating unrelated variables. A karnaugh map reduces the need for extensive calculation by taking advantage of the humanś pattern recognition capability. This also permitts the rapid identification and elimination of potential race conditions. A Karnaugh map is 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 3 and Figure 2.

Table 3: Truth table for two variables

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

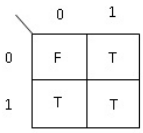


Figure 2: General case of a two variables in k-map

If the number of terms *n* is even then matrix of size $2^{n/2} \times 2^{n/2}$ is created and if the number of terms *n* is odd then a matrix of size$2^{(n-1)/2} \times 2^{(n-1)/2}$ created.
In this research study, the k-map approach on uncertain textual data to find a frequent termset which reduces the database scans and improves the efficiency and accuracy of algorithm.

2 Literature Review

There are number of different approaches available in the literature for frequent pattern mining from uncertain data [1], [10], [11], [13], [14], [20], [21]. This section provide some background and discuss work related to data uncertainty. Some researchers have extended association rule mining techniques to imprecise or uncertain data. They have proposed different approaches and framework.

Leung, et. al. proposed efficient algorithms for the mining of constrained frequent patterns from uncertain data [8] in 2009. They proposed, using U-FPS algorithms, to find the frequent patterns for efficient mining that satisfy the user-specified constraints from uncertain data.

Aggarwal, et. al. proposed a framework for clustering uncertain data streams [9] in 2008. They provide a method for clustering. They use a general model of the uncertainty in which they assume that only a few statistical measures of the uncertainty are available.

Chui, et. al. proposed mining a frequent itemset from uncertain data [10] in 2007. They proposed the U-Apriori algorithm, which was 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 a framework for mining frequent itemsets from uncertain data. A data trimming framework proposed to improve mining efficiency. Through extensive experiments, the data trimming technique can achieve significant savings in both CPU cost and I/O cost.

Aggrawal, et. al. proposed frequent pattern mining with uncertain data [11] in 2009. 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.

Abd-Elmegid, et. al. proposed vertical mining of frequent patterns from uncertain data [13] in 2011. They extended the state-of-art vertical mining algorithm, Eclat, for mining frequent patterns from uncertain data and generated the 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 a comparative study between the proposed algorithm and well known algorithms.

Tang, et. al. proposed mining probabilistic frequent closed itemsets in uncertain databases [14] in 2011. In this paper they pioneer in defining probabilistic frequent closed itemsets in uncertain data. They proposed a probabilistic frequent closed itemset mining (PFCIM) algorithm to mine from uncertain databases.

Ngai, et. al. proposed efficient clustering of uncertain data [22] in 2006. 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 was independent of each other and can be combined to achieve an even higher pruning effectiveness.

Leung, et. al. proposed the efficient mining of frequent patterns from uncertain data [23] in 2007. In this paper they proposed a tree-based mining algorithm (UFP-growth) to efficiently find frequent patterns from uncertain data, where each item in the transactions is associated with an existential probability. They plan to investigate ways to further reduce the tree size.

We briefly describe our basic approach to the problem and then produce the best results. In this paper, uncertain textual data is used to generate the frequent patterns.

3 Preliminary and Problem Definitions

Definition 1. A termset is frequent if and only if its support value is greater than a predefined minimum support threshold ($min\_support$), where *min_support* is user-specified.

Definition 2. The threshold probability of a term is $DP(T_1)$ and threshold probability of another term is $DP(T_2)$,similarity threshold probability of the term *m* is $DP(T_m)$. Then the total threshold probability of the *m* term is $DP(T_1, T_2.....T_m) = DP(T_1) * DP(T_2)..... * DP(T_m) \forall$ *m* terms that are independent.

Definition 3. Support count can be calculated using the following formulas: To calculate *1-Termset*

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

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

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

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

To calculate *m-Termset*

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

$Supp(T_m T_1) = \{A_{ij} \forall A_{ij}$ (container rows of T_m *container column of T_1)/n} (6)

Where *r* and *c* denote the rows and columns of the k-map respectively and R_{set} and C_{set} are the collection of all rows and columns. Table 4 shows the number of possible termsets in the order of terms.

Lemma 1: Two events are independent if the occurrence of one of the events gives us no information about whether or not the other event will occur; that is, the events have no influence on each other. In probability theory said that two events, A and B, are independent if the probability that they both occur is equal to the product of the probabilities of the two individual events, [24] i.e.

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

4 Proposed Method

Our proposed k-map approach that mines from uncertain textual data and finds the frequent termset, satisfies the user-specified succinct constraint. In 4.1, introduce the proposed algorithm that is used in the mining process. In 4.2, explain in detail the algorithm through an illustrative example.

4.1 Proposed Algorithm

Algorithm

START

Input: *T* $= \{T_1, T_2 ..... T_m\}$// set of all termset
D $= \{D_1, D_2 ..... D_n\}$// 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 *1-termset*.
2. for(x=2 to m)
3. {
4. Calculate all combination of *m-termset* of all term *T* and calculate probability for all.
5. Compare the probability of all termset with *DP(T)*.

6. If probability of termset is greater than or equal to the threshold probability then increase the k-map block.
7. $\}$
8. Calculate the support count for *m-termset*.
9. Now compare the support count of termset if $Supp >=$ *min_support* then termset is frequent.

END

4.2 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 $D = \{D_1, D_2, D_3.....D_{20}\}$.

In the above uncertain textual data which is shown in Table 5, 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 their probability set, which is shown in Table 6 and in same way extended for all combination of *3-termset* and find their probability set, which is shown in Table 7.

Now find all combination of *4-termset* and find there probability set, which is shown in Table 8.

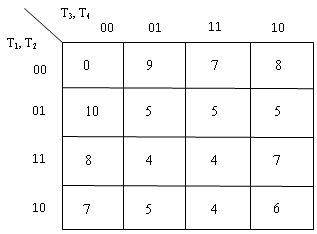
Then, calculate all possibility of termset and then expected threshold based on Lemma 1 as given below.

- Compare the probability of all 1- termset with expected threshold value.
  $\{T_1, T_2, T_3, ...T_m\} \geq \{0.80\}$

- Compare the probability of all 2- termset with expected threshold value.
  $\{\{T_1, T_2\}, \{T_1, T_3\}, \{T_3, T_4\} ..... \{T_1, T_m\}\} \geq 0.64 \{0.80 * 0.80\}$
- Compare the probability of all 3- termset with expected threshold value.
  $\{\{T_1, T_2, T_3\}, \{T_1, T_2, T_4\}, \{T_2, T_3, T_4\} ... \{T_2, T_3, T_m\}\} \geq 0.51 \{0.80 * 0.80 * 0.80\}$
- Compare the probability of all 4- termset with expected threshold value.
  $\{\{T_1, T_2, T_3, T_4\} ..... \{T_1, T_2, T_3, T_m\}\} \geq 0.41 \{0.80 * 0.80 * 0.80 * 0.80\}$
- Compare the probability of all m- termset with expected threshold value.
  $\{\{T_1, T_2, T_3, T_4, ...\} ..... \{T_1, T_2, T_3, , ...T_m\}\} \geq DP_m \{0.80 * 0.80 * 0.80 * 0.80 * .... * 0.80\}$

Let $T = (T_1, T_2, T_3, T_4)$ be the set of all terms, where each term has a probabilistic value . if its value is greater than expected threshold then it can hold 1 either 0 . $D = (D_1, D_2, D_3, ..., D_n)$ the set of all documents. Each documents D_n contains a subset of terms chosen from *T*. In 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 Figure 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 show the frequencies of the items.

First, calculate the support count of the *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 $min\_support = 1$.



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

Figure 3: k-map of term frequency

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

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* .

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

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

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

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

As seen above the *2-frequent termsets* $\{T_1, T_2\}, \{T_1, T_3\}, \{T_2, T_3\}, \{T_3, T_4\}$ are frequent termsets and $\{T_1, T_4\}, \{T_2, T_4\}$are not frequent because their support value is less than the defined 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 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.

5 Experimental Evaluation

In this section, the performance of proposed approaches compared with the existing classical frequent pattern mining algorithms of UApriori [1], UH-mine [6], and UFP-growth [3]. We compare our algorithm and their performance with the state-of-the-art frequent itemset mining algorithm for uncertain textual data sets. In all cases the performance of our algorithm found to be encouraging because this approach requires only one pass of database scan. Thus it reduces I/O time as well as CPU time. UApriori required n-1 (where n is number of items) time scanning of database and UFP-growth required two times scan of the database to generate frequent patterns. 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++.

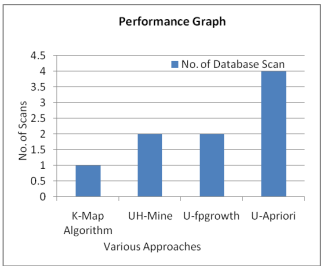


Figure 4: Performance comparison with traditional approaches

Datasets: To test the proposed approach, two different kinds of popular datasets used: 20 News group [8], [25] and Reuters [8], [25], 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. The detailed information of these datasets is described as follows:
20 Newsgroup: 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.

6 CONCLUSIONS

Several data mining methods have been proposed in the last few decades. This paper study the problem of finding frequent termsets from uncertain textual data. The proposed karnaugh map technique works well for mining patterns (termset) from uncertain data. We have introduced a new approach UTDKM for frequent termset, and we have also compared the performance of this algorithm with the already existing algorithms like UApriori, UH-mine, and UFP-growth. It requires only a single database scan and stores the data in the form of frequency in a karnaugh map. This 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 performance depreciates exponentially with the increase in number of variables. Further this work can be extended by the use of QuineMcCluskey algorithm, which works well with a high number of variables so it will solve the problem of high dimensional data, and the frequent termset may used for document clustering.

.

Acknowledgment

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

Competing Interests

The authors declare that no competing interests exist.

References

[1] Aggarwal C C., An Introduction to uncertain data algorithm and applications, Advances in Database Systems. 2009; 35; 1–8.

[2] Rajput D S., Thakur R S., Thakur G S., Rule Generation from Textual Data by using Graph Based Approach, International Journal of Computer Application (IJCA). 2011; 31(9); 36–43.

[3] Han I., Kamber M., Data Mining concepts and Techniques, M. K. Publishers. 2000; 335–389.

[4] Rajput D S., Thakur R S., Thakur G S., Fuzzy Association Rule Mining based Frequent Pattern Extraction from Uncertain Data, IEEE 2nd World Congress on Information and Communication Technologies (WICT'12). 2012; 709–714.

[5] Thakur R S., Jain R C., Pardasani K R. Graph Theoretic Based Algorithm for Mining Frequent Patterns, IEEE World Congress on Computational Intelligence Hong Kong. 2008; 629–633.

[6] Agrawal R., Srikant R.titFast algorithms for mining association rules In Proc. VLDB 1994, pp. 487–499.

[7] Rajput D S., Thakur R S., Thakur G S. Fuzzy Association Rule Mining based Knowledge Extraction in Large Textual Dataset, International Conference on Computer Engineering Mathematical Sciences(ICCEMS'12). 2012; 191–194.

[8] Leung C K S., Hao B., Efficient algorithms for mining constrained frequent patterns from uncertain data, Proceedings of the 1st ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data. 2009; 9-18.

[9] Aggarwal C C., Philip S Yu., A Framework for Clustering Uncertain Data Streams, Data Engineering, IEEE 24th International Conference on ICDE'08. 2008; 150-159.

[10] Chui C K., Kao B., Hung E., Mining Frequent Itemsets from Uncertain Data, Springer-Verlag Berlin Heidelberg PAKDD'07. 2007; 4426; 47-58.

[11] Aggarwal C C., Yan L., Wang Jianyong, Wang Jing., Frequent pattern mining with uncertain data, In Proc. KDD. 2009; 29-37.

[12] Leung C K S., Carmichael C L., Hao B., Efficient mining of frequent patterns from uncertain data, In Proc. IEEE ICDM Workshops'07. 2007; 489-494.

[13] Abd-Elmegid L A., El-Sharkawi M E., El-Fangary L M., Helmy Y K., Vertical Mining of Frequent Patterns from Uncertain Data, Computer and Information Science. 2010; 3(2); 171–179.

[14] Tang P., Peterson E A., Mining Probabilistic Frequent Closed Itemsets in Uncertain Databases, 49th ACM Southeast Conference.2011; 86-91.

[15] Deshpande A., Guestrin C., Madden S R., Hellerstein J M., W. Hong., Model-Driven Data Acquisition in Sensor Networks, VLDB; 2004.

[16] Chen H., Ku W S., Wang H., Sun M T., Leveraging Spatio-Temporal Redundancy for RFID Data Cleansing, In SIGMOD. 2010.

[17] Pelekis N., Kopanakis I., Kotsifakos E E., Frentzos E., Theodoridis Y., Clustering Uncertain Trajectories, Knowledge and Information Systems. 2010.

[18] Khare N., Adlakha N., Pardasani K R., Karnaugh Map Model for Mining Association Rules in Large Databases, International Journal of Computer and Network Security. 2009; 1(2); 16–21.

[19] Lin Y C., Hung C M., Huang Y M., Mining Ensemble Association Rules by Karnaugh Map, World Congress on Computer Science and Information Engineering. 2009; 320–324.

[20] Zhang Q., Li F., Yi K., Finding frequent items in probabilistic data, In Proc. ACM SIGMOD'08. 2008; 819–832.

[21] Appell D., The New Uncertainty Principle, Scientific American; 2001.

[22] Ngai W K., Kao B. , Chui C K., Efficient Clustering of Uncertain Data, Proceedings of the Sixth International Conference on Data Mining (ICDM'06), 2006; 2701–2709.

[23] Leung C K S., Carmichael C L., Hao B., Efficient mining of frequent patterns from uncertain data, In Proc. IEEE ICDM Workshops. 2007; 489-494.

[24] http://www.stats.gla.ac.uk/steps/glossary/probability.html#probability

[25] Huang J., Antova L., Koch C., Olteanu D. MayBMS: A probabilistic database management system, in Proc. ACM SIGMOD'09. 2009; 1071–1074.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
www.sciencedomain.org/review-history.php?iid=313&id=6&aid=2428

Table 4: All possible frequent termsets with m terms

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

Table 5: 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

Table 6: 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 7: 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

Table 8: 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