International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 7 (2015) pp. 16197-16217

http://www.ripublication.com

An Efficient Text Pattern Mining and Clustering (TPMC) Approach for Record Retrieval

[1]A.Rajesh Kumar and [2]Dr.R.Sasikala

[1]Associate Professor Department of Computer Science and Engineering Karur College of Engineering
rajeshkumara0280@gmail.com
[2]Professor and Head Department of Information Technology K.S.R College of Technology

Abstract

Knowledge Discovery and Data Mining is one of the recent emerging fields for mining useful information and knowledge from the digital data, which have rapid growth. Even many techniques are available in data mining for the retrieval record based on text pattern mining, still updating discovered patterns is a trouble in datamining. In order to tackle the problems, a record retrieval method using clustering based on text pattern mining is proposed with four stages. In addition with this, this proposed method is performed on the records, based on the two main phases, which are training and testing phases. In the training phase, closed item sets from each record is extracted using support values, followed by the identification of Normalized D-Patterns of records. Noise Negative records are also consider in the proposed work in order to accomplish updated item sets for each records, which leads to the reduction in error making items. After getting accurate updated records, the weight for every record are computed to move for further processes. And then in testing phase, the records which have similar or nearly similar weights are clustered using K-Means clustering. Each record is assigned with a rank, based on the weight of records and the centroid value of clusters in K-Means. As a result, top ranked records are retrieved from the testing phase. Our proposed work is implemented in Java Platform over real-world datasets and the performance of our work is evaluated using the performance measures such as Precision, Recall and F-measure, which shows that our proposed TPMC approach is best for the record retrievals.

Keywords: *Data Mining, Text Pattern Mining, Closed Item sets, Normalized D-Patterns, Noise Negative Records, K-Means Clustering*

1.Introduction

Text mining, also known as text data mining [3] or knowledge discovery from textual databases, refers generally to the process of extracting interesting and non-trivial patterns or

knowledge from unstructured text documents. It can be viewed as an extension of data mining or knowledge discovery from (structured) databases [1, 4]. As the most natural form of storing information is text, text mining is believed to have a commercial potential higher than that of data mining. In fact, a recent study indicated that 80% of a company's information is contained in text documents. Text mining, however, is also a much more complex task (than data mining) as it involves dealing with text data that are inherently unstructured and fuzzy. Text mining is a multidisciplinary field, involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining.

Text mining can be visualized as consisting of two phases: Text refining that transforms free-form text documents into a chosen initemediate form, and knowledge distillation that deduces patterns or knowledge from the initemediate form [2]. Initemediate form (IF) can be semi-structured such as the conceptual graph representation, or which may be structured such as the relational data representation. Initemediate form can be document-based wherein each entity represents a document, or concept based wherein each entity represents an object or concept of interests in a specific domain. The keyword-based approach of Text Mining which is typically used in this field, the pattern-based model containing frequent sequential patterns is employed to perform the same concept of tasks [6]. However, how to effectively use these discovered patterns is still a big challenge. The performance of the pattern deploying algorithms for text mining is investigated on the Reuters dataset RCVI and the results show that the effectiveness is improved by using our proposed pattern refinement approaches [4].

Over the years, people have often held the hypothesis that phrase-based approaches could perform better than the item-based ones, as phrases may carry more "semantics" like information. This hypothesis has not fared too well in the history of Information Retrieval IR [9]. Although phrases are less ambiguous and more discriminative than individual items, the likely reasons for the discouraging performance include: (i) phrases have inferior statistical properties to items, (ii) they have low frequency of occurrence, and (iii) there are large number of redundant and noisy phrases among them [8].

In the presence of these setbacks, sequential patterns used in data mining community have turned out to be a promising alternative to phrases [3], [7] because sequential patterns enjoy good statistical properties like items. To overcome the disadvantages of phrase-based approaches, pattern mining based approaches (or pattern taxonomy models (PTM) [5]) have been proposed, which adopted the concept of closed sequential patterns, and pruned non-closed patterns. These pattern mining based approaches have shown certain extent improvements on the effectiveness. However, the paradox is that people think pattern-based approaches could be a significant alternative, but consequently less significant improvements are made for the effectiveness compared with item-based methods.

There are two fundamental issues regarding the effectiveness of pattern-based approaches: low frequency and misinterpretation [7]. Given a specified topic, a highly frequent pattern (normally a short pattern with large support) is usually a general pattern, or a specific pattern of low frequency [9]. If we decrease the minimum support, there are a lot of noisy patterns would be discovered. Misinterpretation means the measures used in pattern mining (e.g., "support" and "confidence") turn out to be not suitable in using discovered patterns to answer what users want. Knowledge discovery in databases (Fayyad et al. 1996) is an iterative

process of searching for valuable information in large volumes of data in a cooperative effort of humans and computers: humans select the data to be explored, define analysis problems, set goals and interpret the results, while computers search through the data, looking for models and patterns that meet the human-defined goals. The central step in this process is data mining (Witten and Frank 2005), the purpose of which is to automatically build classification models or find descriptive patterns in large data collections. A variant of data mining is text mining (Feldman and Sanger 2007) where models and patterns are extracted from collections of text documents. Text mining is relevant to linguistic research thanks to its ability to (i) process large amounts of text, which is hard to do by hand, and (ii) automatically uncover non-obvious and unexpected patterns in language use, for example in newspaper discourse.

2. Related Works

Recently text pattern mining is one of the emerged topics among the researches. Some of the reviews about text pattern mining are given below.

DnyaneshRajpathak*et al.* [5] have proposed a novel association and text mining system for knowledge discovery (ASTEK) from the warranty and service data in the automotive domain. The complex architecture of modern vehicles makes fault diagnosis and isolation a non-trivial task [2]. The association mining isolates anomaly cases from the millions of service and claims records. On an average, the analysis time is reduced from few weeks to few minutes, which in real life industry are significant improvements.

CoesemansRoel*et al.* [9] have proposed an approach that differs from the text mining methods in two respects. Firstly, none of the above approaches focuses on qualitative discourse analysis of the results. They provide lists of words or word types that are indicative for a certain type of ideologically biased discourse, e.g. racist/non-racist [11], but do not provide the interpretation of features from a discourse-oriented theory. Secondly, these approaches do not take into consideration different combinations of words.

Balahur and Steinberger [7] have explored sentiment analysis in newspaper texts, aiming at discovering the positive or negative opinions expressed in the articles on a given topic. In case of newspapers, they argued, three different components were to be distinguished: The author, the reader and the text itself. Concentrating on analyses of quoted text, they established guidelines for positive and negative sentiment annotation.

An impressive application of statistical and machine learning approaches used in daily monitoring of news from different media [10] is the Europe Media Monitor1, a research and development effort of the European Joint Research Center in Ispra, Italy, that gathers reports from news portals world-wide in 43 languages, classifies the articles, analyzes the news texts by extracting information from them, aggregates the information, issues alerts, and produces visual representations of the information found.

Lin *et al.* [12] have proposed a statistical model for ideological discourse, based on the hypothesis that ideological perspectives can be detected through lexical variations which aimed at contributing to mutual understanding between Palestinians and Israelis through the open exchange of ideas and observed that some words in discourse were used more frequently because of their relation to the text topic, while other words were used more frequently because of the author's particular ideological perspective.

It is not easy to obtain the right information from the Web for a particular Web user or a group of users due to the obstacle of automatically acquiring Web user profiles. The current techniques do not provide satisfactory structures for mining Web user profiles. In order to tackle this problem, Yuefeng Li and NingZhong [13] have proposed a novel approach. The objective of the approach was to automatically discover ontologies from data sets in order to build complete concept models for Web user information needs. They have also introduced a method for capturing evolving patterns to refine discovered ontologies. In addition, the process of assessing relevance in ontology was established. Their work has provided both theoretical and experimental evaluations for the approach. The experimental results have shown that all objectives they expected for the approach were achievable.

3. Problem Definition

Pattern mining plays an important role in many applications, such as bioinformatics and consumer behavior analysis. However, the classic frequency-based framework often leads to many patterns being identified, most of which are not informative enough for business decision-making. In frequent pattern mining, a recent effort has been to incorporate utility into the pattern selection framework, so that high utility (frequent or infrequent) patterns are mined which address typical business concerns such as dollar value associated with each pattern. Extracting useful patterns from the conflict data is particularly challenging because it is longitudinal, sparse and heterogeneous.

4. Proposed TPMC approach for Record Retrieval

Our proposed text pattern mining and clustering approach comprises of two phases such as Testing and Training phases. The complete process within these phases consists of four stages:

1) Closed item set Extraction Phase
2) Normalized D-Pattern Discovery Phase
3) Noise Negative item Pattern Evolution Phase
4) Clustering based Weight Assignment Phase

The proposed system is illustrated in fig. 1.In the training phase, initially, the frequent item sets from every record are extracted and by subsequently extracting the closed item sets from these extracted frequent item sets based on the support value of each item sets. From the extracted closed item sets, the D-Patterns with its corresponding support value are obtained and this results into Normalized D-Pattern. Then based on the support value of every item sets of the Normalized D-Pattern of the record, the weight are assigned to every record. Then the Noise Negative records are also converted with the format of Normalized D-Pattern. The error making items are rejected, if the Noise Negative record is the complete conflict offender one and the support of error making items are reshuffled, if the record is the partial conflict offender record. Thus, the chance of making errors in the record is reduced by updating new support values.

In the testing phase, the updated item values are weighted with this new value and then the similar weighted record are clustered using K-Means Clustering Algorithm. The matched records are ranked based on the distance between the weights and the centroid in the K-Means Clustering Algorithm that assigned to each record in the same cluster. Top ranking cluster records are retrieved as the result of the proposed work.

4.1. Closed Item Set Extraction Phase

In this paper, we assume that all the records are spitted in to frequent item set and closed item set. Based on the support value of each item sets, frequent item set are extracted from every record. Based on these frequent item sets the closed item set is extracted.

Let T= $\{t_1, t_2, t_3 \ldots\ldots t_k\}$ be a set of items and R be a training set of record, which consists of set of positive record R^+ ; set of negative record R^- . A set of item set is referred to as item set.

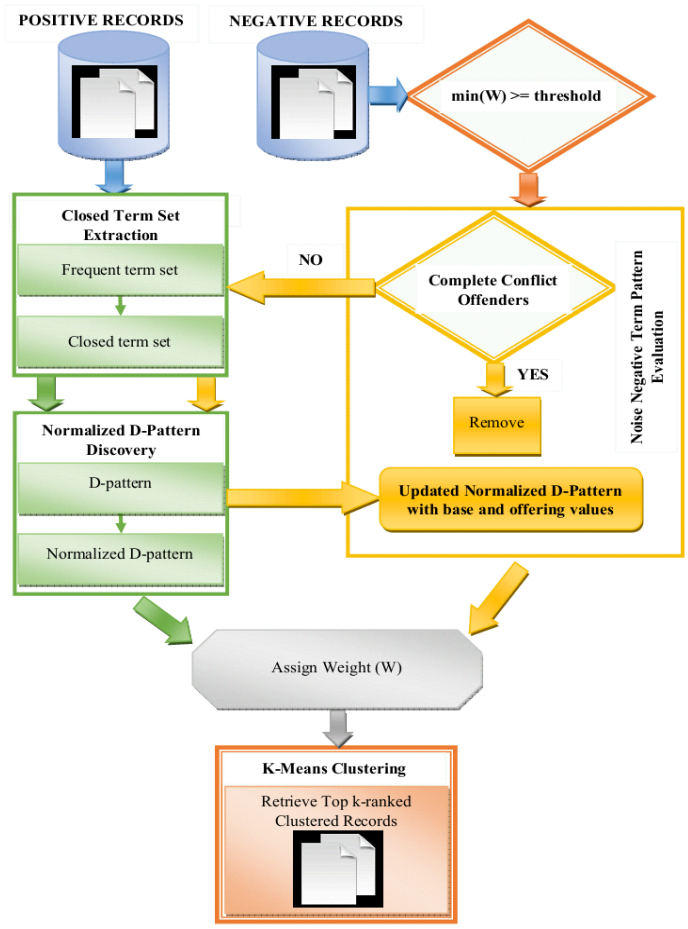

Fig. 1: Proposed TPMC approach for Record Retrieval

4.1.1. Frequent Item set

The concept of frequent item set was first introduced for extractingrecorddatabases. Let $I = \{i_1, i_2 \ldots\ldots i_n\}$ be a set of all items. A k-item set α , which consists of k items from I , is frequent. The frequent patterns are discovered based on the records, the item set the frequent and closed patterns are record based on the records, x the item set in records, X is used to denote the covering set of x for d, which includes all file $rf \in ps(R)$ such that

$$x \subseteq rf \tag{1}$$

$$X = \{rf \mid rf \in (r) X \subseteq rf\} \tag{2}$$

Its absolute support is the number of occurrences of X in $fs(r)$

$$Sup_a(x) = |\mathrm{X}\ | \tag{3}$$

Its relative support is the fraction of the file that contains the pattern, which is given in the following eqn. (4).

$$Sup_r(x) = \frac{|A\ |}{|\ fs(r)\ |} \tag{4}$$

A item-set x is called frequent pattern if it's $\sup_r (or \sup_a) \geq \min\_\sup$ a minimum support. Lists a set of files for a given record d,where $fs(r) = \{rf_1 rf_{2,} rf_3 .....rf_6\}$ and duplicate items were removed. A major challenge in extracting frequent pattern from a large data set is the fact that such extraction often generates a huge number of patterns satisfying the *min_sup* threshold, especially when *min_sup* is set low. This is because if a pattern is frequent each of its sub patterns is frequent as well. A large pattern will contain an exponential number of smaller, frequent sub patterns. Patterns can be structured into taxonomy by using the subset relation. Smaller patterns in the taxonomy are usually more general because they could be used frequently in both positive and negative records and larger patterns are usually more specific since they may be used only in positive records. Thesemantic information will be used in the pattern taxonomy to improve the performance of using closed patterns in text mining.

4.1.2. Closed Item set

A sequential pattern X is called frequent pattern if its relative support (or absolute support) is a minimum support. Some property of closed patterns can be used to define closed sequential patterns.

Given a item set x, its covering set is a subset of records. Similarly, given a set of records $Y \subseteq fs(r)$, we can define its item set, which satisfies $termset(Y) = \{i | \forall rf \in Y \Rightarrow i \in rf\}$.The closure of X is defined as follows in eqn. (5),

$$cls(x) = termse(X) \tag{5}$$

A pattern x (also a item set) is called closed, if and only $x = cls(X)$.

Let X be a closed pattern. We can prove that

$$\sup_a(x_1) < \sup_a(x) \tag{6}$$

For all patterns $x_1 \supset x$; otherwise, if $\sup_a(x1) = \sup_a(x)$, $X_1 = X$, where $\sup_a(x_1)$ *and* $\sup_a(x)$ are the absolute support of pattern x_1 and x, respectively. We also have

$$cls(x) = termset(X) = termset(X_1) \supseteq x_1 \supset x, \text{(i.e.,)}\ cls(x) \neq x \quad (7)$$

Let *min _sup* = 50%, using the above definitions. Table I illustrates the frequent itemsets and their covering sets.

Table I: Frequent Item sets and Covering Set

Records	rf_1	rf_2	rf_3	rf_4	rf_5	rf_6	-	-	-
Items	i_1, i_2	i_3, i_4, i_6	i_3, i_4, i_5, i_6	i_3, i_4, i_5, i_6	i_1, i_2, i_6, i_7	i_1, i_2, i_6, i_7	-	-	-
Frequent item sets	i_3, i_4,	i_3, i_4	i_3, i_6	i_4, i_6	i_3	i_4	i_1, i_2	i_2	i_6
Covering sets	rf_2, rf_3, rf_4	rf_2, rf_3, rf_4	rf_3, rf_4	rf_2, rf_3, rf_4	rf_2, rf_3, rf_4	rf_2, rf_3, rf_4	rf_1, rf_5, rf_6	rf_1, rf_5, rf_6	rf_2, rf_3, rf_4, rf_5, rf_6

Not all frequent patterns in Table 1 are useful. For example, pattern $\{i_3, i_4\}$, always occurs with item i_6 in records, i.e., the shorter pattern, $\{i_3, i_4\}$ is always a part of the larger pattern $\{i_3, i_4, i_6\}$, in all of the records. Hence, we believe that the shorter one, $\{i_3, i_4\}$, is a noise pattern and expect to keep the larger pattern, $\{i_3, i_4, i_6\}$, only.

4.1.3 Pattern Taxonomy Model (PTM)

Patterns can be structured into taxonomy by using the subset relation. In PTM method, each record 'r' are spited into files 'f', which yields $fs(r)$. From the set of record in the frequent patterns and the covering sets are discovered for each. Smaller patterns in the taxonomy, patterns are usually more general because they could be used frequently in both positive and negative records. Larger patterns, in the taxonomy are usually more specific since they may be used only in positive records.

For the example of Table I, where we have illustrated a set of records, and the discovered frequent patterns in Table I if assuming min sup = 50%. There are, however, only three closed pattern in this example. They are $< i_3, i_4, i_6 >, < i_1, i_2 >,$ and $< i_6 >$

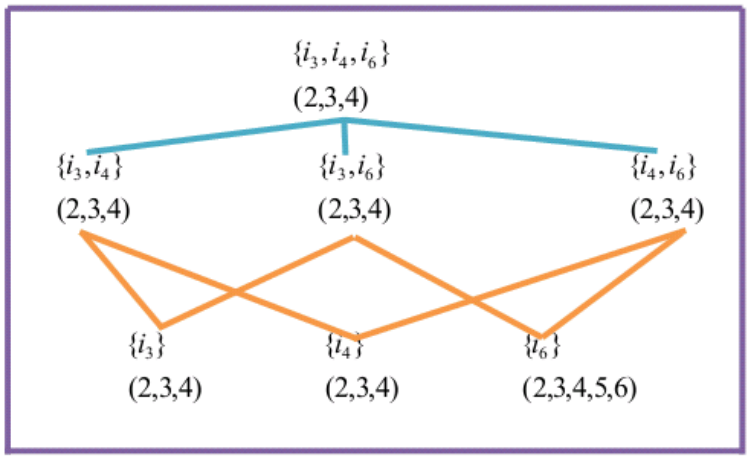

Fig.2: Example for Pattern taxonomy model

Fig.2 illustrates an example of the pattern taxonomy for the frequent patterns in Table 1, where the nodes represent frequent patterns and their covering sets; non-closed patterns can be pruned; the edges are "is-a" relation. After pruning, some direct "is-a" retaliations may be changed, for example, pattern $\{i_6\}$ would become a direct sub pattern of $\{i_3, i_4, i_6\}$ after pruning non-closed patterns. Smaller patterns in the taxonomy, for example pattern $\{i_6\}$, (see Fig. 2) are usually more general because they could be used frequently in both positive and negative records; and larger patterns, for example pattern $\{i_3, i_4, i_6\}$, in the taxonomy are usually more specific since they may be used only in positive records. The semantic information will be used in the pattern taxonomy to improve the performance of using closed patterns in text mining.

4.2. Normalized D-Pattern Discovery Phase

From the extracted closed item sets, the D-Patterns with its corresponding support value are obtained.Then the result of this D-Pattern is move for further process to make normalized D-pattern. After the weight are assigned to every record based on the support value of every item set of the normalized D-Pattern of the record. Then the noise negative records are also converted with the format of Normalized D-Pattern

4.2.1. D-Pattern Representation

In order to use the semantic information in the pattern taxonomy to improve the performance of closed patterns in text mining, we need to interpret discovered patterns by summarizing them as d-patterns in order to accurately evaluate item weights (supports). The rational behind this motivation is that d-patterns include more semantic meaning than items that are selected based on a item-based technique $(e.g., if * irf)$. As a result, a item with a higher $it * irt$ value could be meaningless if it has not cited by some d-patterns (some important parts in records). The evaluation of item weights (supports) is different to the normal item-based approaches. In the item-based approaches, the evaluation of item weights is based on the distribution of items in records. In this research, items are weighted according to their appearances in discovered closed patterns.

D-Pattern

D-pattern mining algorithm is used to discover the D-patterns from the set of records. The efficiency of the pattern taxonomy mining is improved by proposing and to finding all the closed sequential patterns, which is used as the well-known appropriate property in order to reduce the searching space. It describes the training process of finding the set of d-patterns for every positive record. The main focus is the deploying process, which consists of the d-pattern discovery and item support evaluation. All discovered patterns in a positive records are composed into a d-pattern giving rise to a set of d-patterns. Thereafter, item supports are calculated based on the normal forms for all items in d-patterns.

4.2.2. Normalized D-Pattern

Let RF be a set of d -patterns in R^+, and $f \in rf$ be a d-pattern. We call $f(i)$ the absolute support of item i, which is the number of patterns that contain i in the corresponding patterns taxonomies. Let w be assigning weight for incoming record r. In order to effectively deploy patterns in different taxonomies from the different positive records, d-patterns will be normalized using the following assignment sentence:

$$\beta(f_i) = \{(i_1, w_1), (i_2, w_2), ..., (i_k, w_k)\} \tag{8}$$

For all $f_i \in rf$,where, $F_i = \{(i_1, t_1), (i_2, t_2), ..., (i_k, t_k)\} \in rf$

$$\text{And,} \quad w_i = \frac{f_i}{\sum_{j=1}^{k} f_j} \tag{9}$$

To assign weight for all incoming record r based on their corresponding weight $'W'$ functions the following formulae is used,

$$Weight(r) = \sum_{i \in I} w(i)\tau(i, r) \tag{10}$$

4.3. Noise Negative Item Pattern Evolution Phase

In this section we discuss about noise negative item pattern evolution. A noise negative record nr in R^- is a negative records that the system falsely identified as a positive, that is $weight(nr) = Threshold(rf)$. In order to reduce the noise, We need to track which d-pattern have been used to give rise to such an error. We call these patterns offenders of nr.

4.3.1 Threshold

A threshold is usually used to classify records in to relevant or irrelevant categories. Using the d-patterns, the threshold can be defined naturally as follows:

$$Threshold(RF) = \min_{r \in RF} \left(\sum_{(i,w) \in \beta(p)} Support(i) \right) \tag{11}$$

4.3.2. Offenders

An offenders of nr is a d-pattern that has at least one item in nr. The set of offenders of nr is define as

$$\Delta(nr) = \{r \in RF \mid termset(p) \cap nd \neq \phi\} \tag{12}$$

There are two types of offenders. They are given below

I. A Complete conflict offenders of nr

II. A partial conflict offenders which contains part of nr

The basic idea of updating patterns is explained as follows: complete conflict offenders are removed from d-patterns first. For partial conflict offenders, their item supports are reshuffled in order to reduce the effects of noise records. The basic idea of updating patterns is explained as followers initially remove all complete conflict offenders from d-patterns for partial conflict offenders, if the items support are reshuffled in order to reduce the effect of noise records.

Table II: Algorithm for Updated Normalized D-Pattern Finding in Noise Negative Documents

Input: A training set $R = R^+ \bigcup R^-$; a set of d-pattern FR; and an experimental coefficient μ **Output:** A set item support pairs nr.

$nr \leftarrow \phi$;

$T \leftarrow threshold$;

$T = Threshold(rl)$

For each $noise\ negative\ record\ nr \in R^- do$

If weight $(nr) \geq T$

$then\ \Delta(nr) = \{f \in RF \mid termset(f) \mid nd \neq \phi\}$;

$NRF = \{\beta(f \mid f \in RF\}$;

$Shuffling\ (nr), \Delta(nr), NRF, \mu NRF)$;

$for\ each\ r \in NRF\ do$

$nr \leftarrow nr \oplus f$;

end

end

The input of this algorithm are set of d-patterns DP, a training set $R = R^+ \bigcup R^-$.The output is a composed of d-patterns. Line3 in Algorithm is used to estimate the threshold for finding the noise negative records. Lines4 to 11 revise item supports by using all noise negative records. Line5 is to find noise record and the corresponding offenders. Line6 gets normal form of d-patterns NRF. Line 7 calls algorithm shuffling to update NDP according to noise records. Lines 8 to 10 compose updated normal forms together.

4.3.3. Steps to obtain Updated Normalized D-Patterns in Noise Negative Documents with a simple Example

Input Values: $R = \{(i_1, 3), (i_2, 3), (i_3, 3) \mid (i_4, 3), (i_6, 8)\}$

Normal form: $\{(i_1, 3/20), (i_2, 3/20), (i_3, 3/20), (i_4, 3/20), (i_6, 2/5)\}$

STEP 1: Assume $nr = \{i_1, i_2, i_6, i_9\}, \hat{r}$ it will be a partial conflict offenders

STEP 2: Apply $itemset(\hat{f}) \mid nr = \{i_1, i_2, i_6\} \neq 0$

$$nr = \{i_1, i_2, i_6, i_9\}$$

$$\hat{r} \bigcap nr = \{i_1, i_2, i_6\}$$

STEP 3: $\mu = 2$ shuffling the offender value and support value

$$Offering = (1 - 1/\mu) = x \sum_{i \in (item\,set(f) - nr)} Support(i); \quad (13)$$

$$= 1/2 * 3/20 + 3/20 + 2/5$$

$$= 7/20$$

STEP 4: Shuffling the Base value and support values

$$Base = \sum_{t \in (termset(f) - nr)} Support(i); \quad (14)$$

$$r - nr = \{i_3, i_4\}$$

$$Base = 3/20 + 3/20$$

$$= 3/10$$

Using the algorithm shuffling we can get the updated normal form.

STEP 5:Updating the new support value

for each item i in itemset(p) *do*

if $i \in nr$

then, $Support(i) = \frac{1}{\mu} * Support(i)$ (15)

else

$$Support(i) = Support(i) \times (1 + offering \div base) \quad (16)$$

Based on the eqn. (16),

$$Support = \{(i_1, 3/40), (i_2, 3/40), (i_3, 13/40), (i_4, 13/40), (i_6, 1/5)\} \quad (17)$$

Here, the new support value is updated by reducing the chance of making errors in the record.

4.4. Clustering Based Weight Assignment Phase

Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Here we clustered the updated values based on the weights.

4.4.1. K-Means Clustering Algorithm

K-Means Clustering algorithm is one of the simplest unsupervised learning algorithm that solve the well-knownclustering problem.Clustering techniques have a wide use and importance in nowadays.The procedure follows a simple and easy way to classify a given files through a certain number of clusters. Efficiency of the original k-means algorithm heavily relies on the initial centroids. Initial centroids also have an influence on the number of iterations required while running the original k-means algorithm. The computational complexity of the original k-means algorithm is very high, specifically for massive files. This paper presents an enhanced method for finding the top ranking cluster record.

4.4.2. Steps to K-Means Clustering Algorithm

Input: The no of positive record and Negative Record

Output: A set of clustered records

1.Identifying unique words from the given record

2.Selection of weight for the record

3.Specifying the value of k (number of clusters)

4.Randomly select K record and place one K selected record in each cluster based on its weight

5.Compute centroid for each K-Clusters

6.Compute weightof each records *r* for each K-Clusters

7.Again using weight*Wi* for each record, Find the distance between the centroid and weight*Wi*

8.Now place the record in the cluster based on similarity between record and the centroid of clusters.

9. After placing all the record in the clusters compare the previous iteration clusters

10. If the entire cluster contains same record in previous and current iteration then iteminate the algorithm here and we will be obtaining the top ranking clustered record.

11. Else repeat through step 7.

4.5. Main Phases of Proposed TPMC Approach

In this model includes two phases. They are

1.Training phase

2.Testing phase

4.5.1. Training Phase

In the training phase, the proposed model mainly find d-patterns in positive records R^+ based on a $\min\_\sup$, and evaluates item supports by deploying d-pattern to items, followed by the generation of Normalized D-Patterns. This is attained by extracting frequent patterns followed by closed item patterns. It also calls algorithm in table II, $Alg\,orithm\,(R^+, R^-, RF, \mu)$ to revise item supports using noise negative records in D-based

on an experimental coefficient μ . Finally, the weights for both positive and negative records are assigned.

4.5.2 Testing Phase

In the testing phase, it evaluates weights for all incoming records then can be sorted based on these weights. These updated weights of every record are then clustered using K-Means Algorithm by considering these weights as input and which results into top k-ranked records as per our requirement of records.

Hence, the required records are effectively retrieved with the aid of our proposed Text Pattern Mining and Clustering approach.

5. Results and Discussions

The proposed retrieval of text pattern mining work is implemented in the platform, JAVA. Our proposed work is implemented based on clustering using K-Means algorithm, which gives very fast and accurate retrieved records from our efficiency of work. The efficiency of our work is examined based on various performance metrics and the sample results of the experiments are given in this section.

5.1. Experimental Results of Proposed work

The proposed text pattern mining is worked out with various real time records. Each record is collectively given for our work to get effective text pattern mining of any of the testing record. Among the whole records, some records are utilized for training and some for testing purposes. And also, some of the records are positive records and some are negative records. The negative records are also used in our proposed work for the effective retrieval of records. Sample records are given below in the following fig. 3.

Input Record 1:

The figure based on the first topics group (r101 r150) for PTM (IPE) is less than that based on the other group (r151 r200). This can be explained in that the high proportion of closed patterns is obtained by using PTM (IPE) based on the first topics group.A further investigation in the comparison of PTM (IPE) and TFIDF in top-20 precision on all RCV1 topics is depicted in Fig. 8. It is obvious that PTM (IPE) is superior to TFIDF as it can be seen that positive results distribute over all topics, especially for the first 50 topics. Another observation is the scores on the first 50 topics are better than those on the last fifty. That is because of the different ways of generating these two sets of topics, which has been mentioned before. Vector data is the item used to describe such geometric approximations, constructed using points, line segments, polygons, spheres, cubes, etc.If we consider a relation in which each tuple is a point representing a city or a lake, the above queries can be answered by a join of this relation with itself, where the join condition specifies the distance between two matching tuples.In this section, however, we concentrate on applications in which spatial data plays a central role and in which efficient handling of spatial data is essential for good performance. So when the support was changed, the algorithm which uses frequent sequence tree as the storage structure could find all the sequential patterns without mining the database. For instance, an airplane wing might be modeled as a wire frame using a collection of polygons (that intuitively tile the wire frame surface approximating the wing), and a tubular object may be modeled between two concentric cylinders.

Fig. 3: Sample records for processing our proposed text pattern mining work

The primary phase of our proposed work helps to extract the closed item sets from the record, which is initially attained by the extraction of frequent item sets. The frequent item sets are extracted based on the support value of each items. The following fig. 4 shows the sample results for the process of extracting frequent item sets.

database,results	tuple,mining,results
results,tuple	sequence,observation,mining
tuple,results	observation,mining,sequence
results,data	sequence,database,mining
data,results	database,mining,sequence
order,sequence	mining,sequence,tuple
sequence,order	results,sequence,observation
order,observation	sequence,observation,results
order,tuple	observation,results,sequence
sequence,observation	results,sequence,precision,patterns
observation,sequence	sequence,precision,patterns,results
database,sequence	patterns,results,observation,mining
sequence,tuple	results,observation,mining,patterns
tuple,sequence	results,observation,mining,patterns
precision,observation	observation,mining,patterns,results
precision,data	mining,patterns,results,observation
database,observation	patterns,observation,tuple,mining
observation,tuple	observation,tuple,mining,patterns
tuple,observation	tuple,mining,patterns,observation
observation,data	mining,patterns,observation,tuple
data,observation	mining,results,observation,geometric
database,data	results,observation,geometric,mining
tuple,data	observation,geometric,mining,results
data,tuple	geometric,mining,results,observation
database,mining	observation,tuple,geometric,mining
mining,tuple	tuple,geometric,mining,observation
tuple,patterns,observation	geometric,mining,observation,tuple
mining,results,sequence	order,mining,observation,patterns,results,database
results,sequence,mining	observation,patterns,results,database,order,mining
sequence,mining,results	patterns,results,tuple,data,geometric,mining
mining,results,precision	results,tuple,data,geometric,mining,patterns
results,precision,mining	tuple,data,geometric,mining,patterns,results
precision,mining,results	data,geometric,mining,patterns,results,tuple
mining,results,observation	geometric,mining,patterns,results,tuple,data
results,observation,mining	mining,patterns,results,tuple,data,geometric
observation,mining,results	patterns,results,tuple,data,geometric,observation
mining,results,database	results,tuple,data,geometric,observation,patterns
results,database,mining	tuple,data,geometric,observation,patterns,results
database,mining,results	data,geometric,observation,patterns,results,tuple
mining,results,tuple	geometric,observation,patterns,results,tuple,data
	(etc.,)

Fig. 4: Sample results for the frequent item sets extraction from the records

Some of the sample results for the extraction of frequent item sets are given in fig. 4. From the results of frequent item set extraction step only, we can find nearest frequent sets, which is closed item sets. The results of closed item sets extraction from the record is given below in fig. 5, which is done followed by frequent item sets extraction.

[patterns, precision, database, mining, results, observation]
[patterns, geometric, results, database, mining, observation, sequence]
[patterns, results, sequence, order, geometric, mining, observation]
[patterns, results, sequence, order, database, mining, observation]
[patterns, sequence, precision, mining, geometric, results, observation]
[patterns, geometric, precision, data, tuple, results, mining, observation]
[patterns, results, sequence, order, data, tuple, mining, observation]
[patterns, results, database, data, tuple, mining, observation, sequence]
[patterns, results, tuple, data, geometric, mining, observation, sequence]

Fig. 5: Sample results for closed item sets extraction

From the resultant closed item sets from the record, it is required to find D-Patterns. The procedure to find D-Patterns is explained detailed in Section 4. The sample results for the D-patterns for the records with the support value are given in the following table III.

Table III: D-Patterns with Support value

D-Patterns	[methods, results, patterns, mining, data, order, database, pattern, sequence, proposed, approach]
Support value for corresponding D-Patterns	[4, 9, 9, 9, 4, 3, 4, 9, 7, 5, 3]

After find out the D-Patterns from the collection of records, we need to proceed with the next step as Normalized D-Patterns based on average support. The resultant output of Normalized D-Patterns is given in the following table IV.

Table IV: Normalized D-Patterns with Support value

D-Patterns	[information, achieved, sequential, pattern, mining, marketing,, medical, records,, sales, analysis,, on., research, work,, effective, pattern, discovery, technique, proposed, overcome, low-frequency, misinterpretation, problems, text, mining., results, sequential, pattern, mining, items, bought, order, customer,, items, coming, transactions, [4]., sequence, ordered, itemsets, timestamp., represented, by,, where,, itemset., sequence, items, length, sequence., shows, deploying, approach, concepts, users, significant, concept-based, model, cbm, cbm, pattern, matching, model., phrases, information, superhighway, ubiquitous,, information, processing, rapidly, growing, multi, billion, dollar, industry., experimental, results, analysis, results, items, generated, sequential, patterns,, execution, time, memory, usage, existing, incspan, algorithm., construct, updated, cssf-trie, static, database., then,, database, updated, distributed, sources;, here,, developed, proposed, algorithm, mining, constraint, sequential, patterns, progressive, database., existing, methods, focus, concept, frequency, assumption, sequences, behaviors, change, time., order, efficiently, capturing, dynamic, nature, data, addition, mining, problem,, initially., large, complex, datasets,, users, tools, simplify, tasks, managing, data, extracting, information, timely, fashion., book, kinds, introductory, courses, choosing, topics, appropriately,, two-course, sequence, supplementing, material, advanced, readings, course., cssf, trie, updated, including, updated, sequence., then,, updated, cssf-trie, mine, progressive, cssf-patterns, proposed, trie, pattern, mining, algorithm.] [sequential, patterns, correlation, transactions, association, rule, represents, intra,

	transaction, relationships., promising, results, explained, deploying, method, promising, (hypothesis, h2, supported), solving, misinterpretation, problem, combine, advantages, items, discovered, patterns, concepts., region, data, collection, regions., region, data, stored, database, typically, geometric, approximation, actual, data, object., sequential, pattern, mining, related, association, rule, mining,, events, linked, time, [3]., commercial, gis, systems, arcinfo, wide, today,, object, database, systems, aim, support, gis, applications, well., vector, data, describe, geometric, approximations,, constructed, points,, line, segments,, polygons,, spheres,, cubes,, etc., deploying, approach, weights, items, appearances, discovery, concepts., association, rule, mining,, mining, results, items, brought, frequently,, items, transaction., sequential, pattern, mining, algorithms, address, problem, discovering, existent, frequent, sequences, database, [2]., range, queries, occur, wide, variety, applications,, including, relational, queries,, gis, queries,, cad/cam, queries., performance, rocchio, prob, method, corresponds, finding, [50]., relation, point, representing, city, lake,, queries, answered, relation, itself,, condition, specifies, distance, matching, tuples., section,, however,, concentrate, applications, spatial, data, plays, central, role, efficient, handling, spatial, data, essential, performance., techniques, association, rule, mining,, frequent, itemset, mining,, sequential, pattern, mining,, maximum, pattern, mining,, closed, pattern, mining.]
Support value for corresponding Normalized D-Patterns	[4, 27, 9, 45, 8, 6, 4, 45, 21, 15, 3] [0, 18, 18, 27, 24, 0, 12, 45, 0, 0, 3]

Then, we find whether any records are in negative form (i.e. noise negative records). If the items are in noise negative form, then the errors are removed, if the record items are complete conflict offenders. And also, update the record with its weights, if the record items are partial conflict offenders. Finally, all the positive and negative records are collectively trained and then the input test record is given as the input record. Based on the weights of each record, the top k-ranked records are retrieved. This is done by the use of K-Means algorithm and the results of K-Means with training and testing records are given in following fig. 6.

Clusters	Centroids
[294] [894, 1157, 1557, 1665, 1845] [2182, 2345, 2803, 3000]	[294.0] [1423.6] [2582.5]

(a)

Cluster and Distance values for Trained Records	
Cluster Values for Trained Records	[294.0, 894.0, 1157.0, 1557.0, 1665.0, 1845.0, 2182.0, 2345.0, 2803.0, 3000.0]
Distance for Trained Records	[0.0, 600.0, 863.0, 1263.0, 1371.0, 1551.0, 1888.0, 2051.0, 2509.0, 2706.0]

(b)

Weight for all Records	
Weight for Trained Records	[294.0, 894.0, 1157.0, 1557.0, 1665.0, 1845.0, 2182.0, 2345.0, 2803.0, 3000.0]

Weight of Input Records	[188.00000000000006, 336.00000000000006, 529.0000000000001]

(c)

Correct Output Record
Input Record-1, 188.00000000000006 is Inside 294.0 Cluster Input Record-2, 336.00000000000006 is Inside 1423.6 Cluster Input Record-3, 529.0000000000001 is Inside 1423.6 Cluster
Record-1, 188.00000000000006 is near: 600.0 Distance value➔Near 1 Positive Record Record-2, 336.00000000000006 is near: 600.0 Distance value➔ Near 3 Positive Record Record-3, 529.0000000000001 is near: 600.0 Distance value ➔Near 4 Positive Record

(d)

Fig 6: K-Means Algorithm – (a) Clusters and its corresponding centroids (b)Cluster and Distance values for Trained Records (c)Weight for all Records (d) Accurate Result for the input testing record

Thus, top k-ranked records are selected for any of the given input record as testing phase. Hence, the sample results show the output of our proposed text pattern mining work.

5.2. Evaluation metrics

The effectiveness of our proposed text pattern mining work with K-Means clustering is evaluated with some of the evaluation metrics. An evaluation metric is used to evaluate the effectiveness of record retrieval systems based on texts and to justify theoretical and practical developments of these systems. It consists of a set of measures that follow a common underlying evaluation methodology. Some of the metrics that we have chosen for our evaluation purpose are Recall, Precision and the F-measure.

$$\textbf{Precision},\ P = \frac{|(related\ records) \cap (retrieved\ \ records)|}{|(retrieved\ records)|} \tag{18}$$

$$\textbf{Recall},\ R = \frac{|(related\ records) \cap (retrieved\ records)|}{|(relevant\ records)|} \tag{19}$$

$$\textbf{F- Measure},\ F = \frac{2PR}{(P+R)} \tag{20}$$

As suggested by above equations in the field of record retrieval based text pattern mining,

1. ***Precision*** is the fraction of retrieved records that are relevant to the search,
2. ***Recall*** is the fraction of the records that are relevant to the input record that are successfully retrieved,
3. ***F-measure*** that combines precision and recall is the harmonic mean of precision and recall.

5.3. Performance Evaluation of our proposed work

The performance of our proposed record retrieval based text pattern mining work is evaluated on the basis of the above evaluation metrics such as Precision, Recall and F-Measure. For a

given testing record, the results of related top k records are retrieved. For this process, the performance is evaluated and the results are tabulated in table V.

Table V: Performance evaluation for our proposed work with the metrics Precision, Recall and F-Measure

Lists of Records	Precision (in %)	Recall (in %)	F-Measure (in %)
Record-1	0.88	1	0.93
Record-2	0.9	1	0.94
Record-3	0.875	1	0.93
Record-4	0.8	1	0.888
Record-5	0.75	1	0.8571
Record-6	0.83	1	0.907
Record-7	0.857	1	0.922
Record-8	0.71	1	0.83
Record-9	0.88	1	0.936
Record-10	0.857	1	0.922

The graphical representation based on the table V is given in fig. 7 with various evaluation measures Precision, Recall and F-Measure.

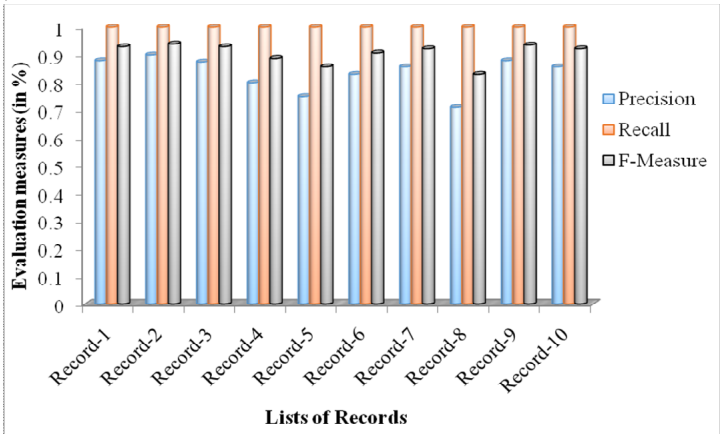

Fig. 7: PerformanceEvaluation graph for the proposed record retrieval work based on text pattern mining

Totally, 10 records are evaluated for the proposed record retrieval work based on text pattern mining with K-means algorithm. Here, these 10 records are also considered as the testing records. Before testing with these set of records, initially we should process with some of the records which are called as training records. Based on the process and results of training records, the results of testing phase are occurred with an effective manner. From the observation of the proposed results, we can say that our proposed work is better suits for the effective record retrieval system with the basis of text pattern mining. The precision values for the 10 records are: 88%, 90%, 87.5%, 80%, 75%, 83%, 85.7%, 71%, 88%, and 85.7%, respectively. And also, the F-measure values for these 10 records are: 93%, 94%, 93%, 88.8%, 85.71%, 90.7%, 92.2%, 83%, 93.6% and 92.2%, respectively. The value of Recall is 100% for all the 10 records. Thus the higher value of these measures indicates that our proposed system of record retrieval is superior one. Moreover, the record number 2 gives

very higher accuracy of retrieval results by providing 90% of Precision and 94% of F-Measure with 100% of Recall values, among the 10 records. On average, 83.39% of Precision and 90.621% of F-Measure values with 100% of Recall value is obtained by our proposed record retrieval system.

5.4. Comparison between our proposed work and existing works

Our proposed work is compared with the existing methods used in the reference number [13] in the literature review part. Text pattern mining approach is compared with the existing work, in which the precision value is mainly compared. The comparison between these two techniques with the measure precision is given in the table VI.

Table VI: Comparison of the metric precision between the existing and proposed techniques

Methods	Average Precision (in %)
Supervised Round2[13]	65.3
Supervised Prob2 [13]	62.2
Proposed method	83.39

The corresponding graph for the table VI is given in the fig. 8, which gives the clear identification of the good performance in our proposed work.

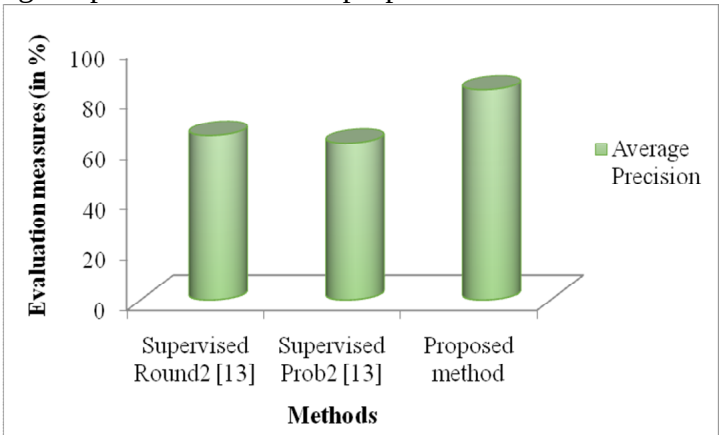

Fig. 8: Comparison between existing and proposed method based on Average Precision

In order to accurately evaluatethe performance of our work, we consider a state-of-art work [13] from the literature review. Yuefeng Li and NingZhong are the authors of the existing work in reference [13] with the research on “Mining Ontology for Automatically Acquiring Web User Information Needs”. They have evaluated the mining results in two ways: (1) Semi-supervised evaluation (2) Supervised evaluation. Here, we are compared the results of Supervised evaluation with two models (Round2 model and Prob2 model), which has provided higher precision value.

In compared with [13], our proposed work provides very superior results of the metric precision. The models Round2 and Prob2 have facilitated 65.3% and 62.2% of precision value, respectively, which is lower value than our proposed work. Because, we are achieved 83.39% of precision value, which is 18.09% and 21.19% higher value than both the models Round2 and Prob2, respectively.Moreover, 100% of recall value is also obtained by our proposed text pattern mining method. Hence, we can prove that our proposed method of text

pattern mining with clustering is the best method by outperforming other existing methods by attaining very good accuracy results.

6. Conclusion

Our Proposed Text Pattern Mining and Clustering methodology has worked with the phases training and testing for the retrieval of records. The evaluation results of our proposed method have shown that our method is the best one byalso updating the additional items from Noise Negative Records for the effective retrieval of records. On average, we have achieved 83.39% of Precision and 90.621% of F-Measure values with 100% of Recall value by using our proposed record retrieval system. Moreover, the comparison was made by an existing work with our proposed system. The analysis from the comparison has also clearly declared that our proposed TPMC for record retrieval is superior method for the text pattern mining by outperforming existing method. Because, we were attained 83.39% of precision value, which is 18.09% and 21.19% higher value than both the models Round2 and Prob2 of existing work, respectively. Hence, the proposed Text Pattern Mining and Clustering approach will be motivate the future researchers to do more research on the field of text pattern mining.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, 1999.
[2] Zhong, Ning, Yuefeng Li, and Sheng-Tang Wu. "Effective pattern discovery for text mining." Knowledge and Data Engineering, Vol. No. 1, pp: 30-44, 2012.
[3] N. Cancedda, E. Gaussier, C. Goutte, and J-M. Renders. Word sequence kernels. Journal of Machine Learning Research, Vol. No. 3, pp: 1059–1082, 2003.
[4] M. F. Caropreso, S. Matwin, and F. Sebastiani. Statistical phrases in automated text categorization. Technical report, Instituto di Elaborazionedell'Informazione, 2000.
[5] Mei, Qiaozhu, and ChengXiangZhai. "Discovering evolutionary theme patterns from text: an exploration of temporal text mining", In Knowledge discovery in data mining, pp: 198-207, 2005.
[6] S. T. Dumais. Improving the retrieval of information from external sources. Behavior Research Methods, Instruments, & Computers, Vol. No. 23, pp: 229–236, 1991.
[7] Weiss, S.M.; Apte, C.; Damerau, F.J.; Johnson, D.E.; Oles, F.J.; Goetz, T.; Hampp, T., "Maximizing text-mining performance," Intelligent Systems and their Applications, IEEE , vol.14, pp :63-69,1999.
[8] Tan, Ah-Hwee. "Text mining: The state of the art and the challenges." In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases, pp: 65-70, 1999.
[9] Weiss, Sholom M., ChidanandApte, Fred J. Damerau, David E. Johnson, Frank J. Oles, Thilo Goetz, and Thomas Hampp. "Maximizing text-mining performance." Intelligent Systems and their Applications, IEEE 14, Vol. No. 4, pp: 63-69, 1999.
[10] Cohen, Aaron M., and William R. Hersh. "A survey of current work in biomedical text mining." Briefings in bioinformatics, Vol. No. 1, pp: 57-71, 2005.
[11] Wu, Sheng-Tang, Yuefeng Li, and YueXu. "Deploying approaches for pattern refinement in text mining." In Data Mining, pp: 1157-1161, 2006.

[12] Mooney, Raymond J., and RazvanBunescu. "Mining knowledge from text using information extraction." Vol. No. 1, pp: 3-10, 2005.
[13] Yuefeng Li and NingZhong, “Mining Ontology for Automatically Acquiring Web User Information Needs”, IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 4, pp. 554-568, April 2006.

A.Rajesh Kumar (Arumugam Rajesh Kumar) obtained his Bachelor’s degree in Computer Science and Engineering from Manonmaniam Sundaranar University. Then he obtained his Master’s degree in Computer Science and Engineering from Anna University of Technology, Tiruchirappalli. Currently, he is an Associate Professor at the Department of Computer Science and Engineering at Karur College of Engineering. His current research interest is Data Mining.

Dr.R.Sasikala received the B.E. degree from Bharathiyar University at Coimbatore, the M.E degree from the Bharathiyar University, and the Ph.D. degree from the Anna University, Chennai. She is currently a Professor and Head of the Department of Information Technology at K.S.R College of Technology. Her research interests are in Mobile Networks, Cloud Computing, Grid Computing and Data Mining. She has published many International Journals, International Conference and National Conference Papers.