International Journal of Research Studies in Computer Science and Engineering (IJRSCSE)
Volume 1, Issue 4, August 2014, PP 53-62
ISSN 2349-4840 (Print) & ISSN 2349-4859 (Online)
www.arcjournals.org

Pattern Deploying and Pattern Evolving Approaches for Text Mining

Mr.K.K.Nikam

Faculty of CSE Department ATS's SBGI Miraj
Shivaji University, Kolhapur
Maharashtra, India
kartikk023@gmail.com

Prof.A.N.Mulla

Faculty of CSE Department ADCET Ashta
Shivaji University, Kolhapur
Maharashtra), India
mulla.anis@gmail.com

Abstract: *Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy (word having multiple meaning) and synonymy (A word having same or nearby same meaning as another word).Instead of keyword based approach which is typically used in this field, pattern based model containing frequent sequential pattern is employed to perform the same concept of task. In this study we propose two approaches based on the use of pattern deploying and pattern evolving strategies. The performance of the pattern deploying algorithms and pattern evolution algorithm for text mining is investigated on the Reuters dataset RCV1 and the results show that the effectiveness is improved by using our proposed pattern refinement approaches*.

Keywords: *Text mining, text classification, pattern mining, pattern deploying, pattern evolving.*

1. INTRODUCTION

Text mining is a method of retrieving useful information from large amount of digital text data. It is therefore crucial that a good text mining model should retrieve the information according to the user requirement. Traditional Information Retrieval (IR) has same objective of automatically retrieving as many relevant documents as possible, whilst filtering out irrelevant documents at the same time. However, IR-based systems do not provide users with what they really need.

Many text mining methods have been developed for retrieving useful information for users. Most text mining methods use Keyword based approaches, whereas others choose the phrase method to construct a text representation for a set of documents. The phrase-based approaches perform better than the keyword-based as it is considered that more information is carried by a phrase than by a single term. New studies have been focusing on finding better text representatives from a Textual data collection. One solution is to use data mining methods, such as sequential pattern mining for Text mining. Such data mining-based methods use concepts of closed sequential patterns and non-closed patterns to decrease the feature set size by removing noisy patterns.

Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from large amount of data. Knowledge discovery can be viewed as the process of nontrivial extraction of information from large databases, information that is implicitly presented in the data, previously unknown and potentially useful for users. Data mining is therefore an essential step in the process of knowledge discovery in databases. With a large number of patterns generated by using data mining approaches, how to effectively use and update these patterns is still an open research issue.

Pattern Discovery Model for the purpose of effectively using discovered patterns is proposed. Proposed system is evaluated the measures of patterns using pattern deploying process as well as finds patterns from the negative training examples using pattern Evolving process.

2. RELATED WORK

In [10], term frequency and inverse document frequency (tfidf) weighting scheme is used for text representation in Rocchio classifiers. In addition to tfidf, the global idf and entropy weighting

scheme is proposed by [5] and improves the performance by an average of 30%. Varying weighting schemes for the bag of words representation approach are given in [7]. The problem of the bag of words approach is how to select a limited number of features among an enormous set of words or terms in order to increase the system's efficiency and avoid the over fitting [15].

To reduce the number of features, many dimensionality reduction approaches have been conducted by the use of feature selection techniques, such as Information Gain, Mutual Information, Chi-Square, Odds ratio, and so on. The details of these selection functions are stated in [15]. The choice of a representation depends on what one regards as the meaningful units of text and the meaningful natural language rules for the combination of these units [15]. With respect to the representation of the content of documents, some research works have used phrases rather than individual words. In [4], combination of unigram and 2-gram is chosen for document indexing in text categorization (TC) and evaluated on a variety of feature selection functions (FEF). Sharma et al. [16] propose a phrase-based text representation for web document management using rule-based Natural Language Processing (NLP) and Context Free Grammar (CFG) techniques. In [1], Alone et al. apply data mining techniques to text analysis by extracting co-occurring terms as descriptive phrases from document collections.

3. REVIEW OF LITERATURE

The main process of text-related machine learning tasks is document indexing, which maps a document into a feature space representing the semantics of the document. Many types of text representations have been proposed in the past. A well-known method for text mining is the bag of words that uses keywords (terms) as elements in the vector of the feature.

Weighting scheme tf*idf (TFIDF) is used for text representation [1]. In addition to TFIDF, entropy weighting scheme is used, which improves performance by an average of30 percent. The problem of bag of word approach is selection of a limited number of features amongst a huge set of words or terms in order to increase the system's efficiency and avoid over fitting. In order to reduce the number of features, many dimensionality reduction approaches are available, such as Information Gain, Mutual Information, Chi-Square, Odds ratio. Some research works have used phrases rather than individual words. Using single words in keyword-based representation pose the semantic ambiguity problem. To solve this problem, the use of multiple words (i.e. phrases) as features therefore is proposed [2, 3]. In general, phrases carry more specific content than single words. For instance, "engine" and "search engine". Another reason for using phrase-based representation is that the simple keyword-based representation of content is usually inadequate because single words are rarely specific enough for accurate discrimination [4]. To identify groups of words that create meaningful phrases is a better method, especially for phrases indicating important concepts in the text. The traditional term clustering methods are used to provide significantly improved text representation.

4. PROPOSED SYSTEM

The Proposed System highlights on new Knowledge discovery model an attempt to effective exploit the discovered patterns in a large data collection using data mining approaches. This model increase efficiency of pattern discovery using different data mining Algorithms with pattern deploying and pattern Evolving method. System use data set from RCV1 (Reuters Corpus Volume 1) which contains training set and test set. Documents in both the set are either positive or negative."Positive "means document is relevant to the topic otherwise "negative". Documents are in XML format. System uses sequential closed frequent patterns as well as non sequential closed pattern for finding concept from data set.

Modules in the proposed system are as follows-

- Data Preprocessing
- Pattern Taxonomy Model/Pattern Discovery
- Pattern deploying
- Pattern Evolving
- Evaluation

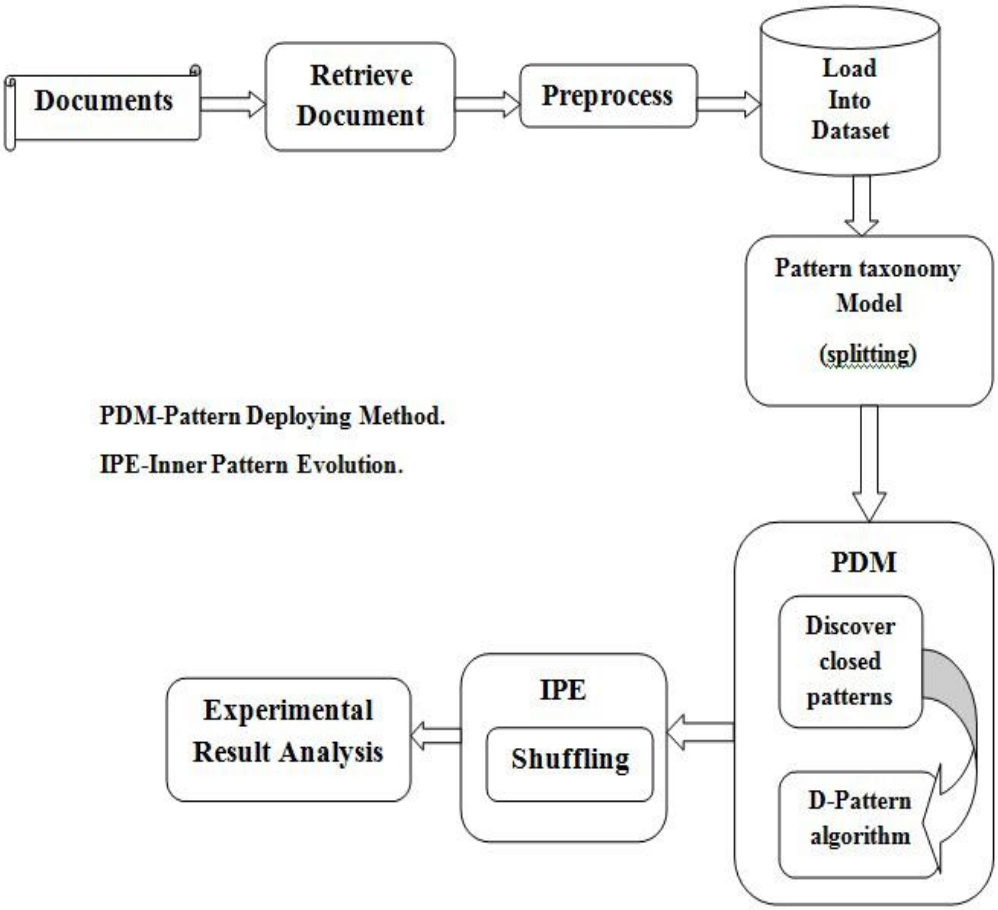

Fig1. *System Flow Diagram*

4.1. Data Preprocessing

This process involves data cleaning and noise removing. It also includes collection required information from selected data fields, providing appropriate strategies for dealing with missing data and accounting for redundant data.

This module consists of following steps

- *Stop Words Removal:* Stop words are those words which are filtered out prior to, or after, processing of natural language data. In this step non informative words removed from document.
- *Text stemming:* Text Stemming is the process for reducing inflected (or sometimes derived) words to their stem base or root form. It generally a written word forms.

4.2. Pattern Taxonomy Model

In this module, the documents are split into paragraphs. Each paragraph is considered to be one document. In each document, the set of terms are extracted. The terms, which can be extracted from set of positive documents

- *Closed Sequential Pattern*: A frequent sequential pattern *P* is a closed sequential pattern if there exist no frequent sequential pattern P' such that P [P' and $supp_a(P) = supp_a(p')$.

4.3. Pattern Deploying

The discovered patterns are summarized in this module. The d-pattern algorithm is used to discover all patterns in positive documents which are then composed. The term support calculates all terms in d-pattern. Term support means weight of the term that is evaluated. These discovered patterns are organized in specific format using pattern deploying method (PDM) and pattern deploying with support (PDS) Algorithms. PDM organizes discovered patterns in <term, frequency> form by combining all discovered pattern vectors. PDS gives same output as PDM with support of each term.

4.4. Pattern Evolving

In this module noisy pattern in the documents are identified. Sometimes, system falsely identifies negative document as a positive documents. That means noise has occurred in positive document. The noisy pattern is named as offender. If positive documents contain the partial offender, the reshuffle process is applied.

4.5. Evaluation of Pattern

This module is regarding evaluation. This compares output of system without deploy and Evolve method with system using deploy and Evolve method. For checking performance of proposed system this module calculates precision, recall and f-measures.

4.6. Implementation

System uses RCV1 documents and retrieves the related information with regard to the training set. Related documents are Preprocessed and transformed into set of transaction based on its nature of document structure. System select pattern discovery algorithm to extract patterns.

Algorithm1. *D-Pattern Mining Algorithm*

Input: positive documents D^+ ; minimum support ,*min_sup*.

Output: d-patterns *DP* ,and supports of terms.

1 $DP = \emptyset$;

2 **foreach** *document* $d \in D^+$ **do**

3 let $PS(d)$ be the set of paragraphs in d;

4 SP =SPMining ($PS(d)$, *min_sup*);

5 $\hat{d} = \emptyset$;

6 **foreach** *pattern* pi $\in SP$ **do**

7 $p=\{(t,1)| t \in p_i\}$;

8 $\hat{d} = \hat{d} \oplus p$;

9 **end**

10 $DP=DP \cup \{\hat{d}\}$;

11 **end**

12 $T=\{t|(t, f) \in p , p \in DP\}$;

13 **foreach** term $t \in T$ **do**

14 $support(t)=0$;

15 **end**

16 **foreach** *d-pattern* $p \in DP$ **do**

17 **foreach** $(t,w) \in \beta(p)$ **do**

18 $support(t) = support(t) + w$;

19 **end**

20 **end**

The pattern taxonomy model improves the semantic meaning of the discovered pattern by using the SPMining, which is helps to reduce the search space. The algorithm 2 describes the training process of finding the set of d-patterns. For every positive document, the SP Mining algorithm is first called giving rise to a set of closed sequential patterns. The main focus is the deploying process, which consists of the d-pattern discovery and word support evaluation. Here words supports are calculated based on the words normal forms for all words in the d-patterns.

The basic definitions of sequences used in this research work are described as follows. Let

T = {t1, t2. . . tk } be a set of all terms, which can be viewed as words or keywords in text documents. A sequence S = s1 , s2 , . . . , sn (si∈ T) is an ordered list of terms. Note that the duplication of terms is allowed in a sequence. An example of sub-sequence: A sequence α = a1 , a2 , . . . , an is a sub- sequence of another sequence β = b1 , b2 , . . . , bm , denoted by α β there exist integers $1 \leq i1 < i2 < \ldots < in \leq m$, such that a1 = bi1 , a2 = bi2 , . . . , an =bin .

Algorithm2. *Sequential closed pattern mining*

Input:-A list of positive documents D+, minimum support(min_sup)

Output:-a set of document vectors Δ

1) Δ=ϕ

2) For each document d in D+ do begin

3) Extract 1 term frequent pattern PL from d

4) SP=SCPM(PL, min_sup) //Algorithm for pattern discovery

5) $\overrightarrow{d}$ =ϕ

6) **For each** pattern p in SP **do begin**

7) $\overrightarrow{d} = \overrightarrow{d} \oplus$ **P'** // p' is expanded from of p

8) End for

9) Δ=Δ ∪ { $\overrightarrow{d}$ }

10) **End for**

After Pattern Deploy, the concept of topic is built by merging pattern of all documents. While the concept is established, the relevance estimation of each document in the test dataset is conducted using the document evaluating equation as shown in (1) in test process. After testing system's performance is evaluated using metrics such as precision, recall and f1-measures shows in equation (2)(3)(4).

Inner pattern evolution shows how to reshuffle supports of terms within normal forms of d-patterns based on negative documents in the training set. The technique will be useful to reduce the side effects of noisy patterns because of the low-frequency problem. This technique is called inner pattern evolution here, because it only changes a pattern's term supports within the pattern. A threshold is usually used to classify documents into relevant or irrelevant categories.

Using the d-patterns, the threshold can be defined in equation (5).A noise negative document nd in D^- is a negative document that the system falsely identified as a positive, that is weight(nd)Threshold (DP). In order to reduce the noise, we need to track which d-patterns have been used to give rise to such an error. We call these patterns offenders of nd. (Offender) .An offender of nd is a d-pattern that has at least one term in nd.

There are two types of offenders:

1) A complete conflict offender which is a subset of nd; and

2) A partial conflict offender which contains part of terms of nd.

The main process of inner pattern evolution is implemented by the algorithm IPEvolving (see Algorithm 3). The inputs of this algorithm are a set of d-patterns DP, a training set D = D+ U D. Here term supports are defined by using all noise negative documents. Which is used to find noise documents and the corresponding offenders and gets normal forms of d- patterns NDP?

Algorithm3. *IPE Evolving*

Input: a training set $D=D^+ \cup D^-$ a set of d-patterns, *DP* ; and an experimental coefficient μ.

Output: a set of term-support pairs *np*.

1 *np*←∅;

2 *threshold= Threshold* (*DP*);

3 **foreach** noise negative documents $nd \in D^{-}$ **do**

4 **if** $weight(nd) \geq threshold$ **then** $\Delta(nd) = \{p \in DP \mid termset(p) \cap nd \neq \emptyset\}$;

5 $NDP = \{\beta(p)\ p \in DP\}$;

6 Shuffling (nd,$\Delta(nd)$, NDP , μ , NDP); //call alg 4.

7 **foreach** $p \in NDP$ **do**

8 $np \leftarrow np \oplus$ p;

9 **end**

10 **end**

Algorithm4. *Shuffling*

Input: a noise document nd, its offenders $\Delta(nd)$,normal forms of d-patterns NDP, and an experimental coefficient μ.

Output: update normal forms of d-patterns NDP.

1 **foreach** *d-patterns p in* Δ (nd) **do**

2 **if** *termset* (p) $\subseteq nd$ **then** $NDP = NDP - \{\beta(p)\}$; //*remove complete conflict offenders*

3 **else** //*partial conflict offender*

4 $offering = (1-1/\mu) \times \sum_{t \in (termset(p) \cap nd)} support\ (t)$;

5 $base = \sum_{t \in (termset(p) - nd)} support\ (t)$;

6 **foreach** *term t in termset* (p) do

7 **if** $t \in nd$ then $support\ (t) = (1/\mu) \times support(t)$; //*shrink*

8 **else** //*grow supports*

9 $support\ (t) = support(t) \times (1 + offering \div base)$;

10 **end**

11 **end**

Documents in dataset are ranked according to their relevance scores. After testing system's performance is evaluated using the metrics such as precision, recall and f-measures. An offender of nd is a d-pattern that has at least one term in nd. The set of offenders of nd is defined in equation (6).

$$Weight(d) = \sum_{t \in TS} support\ (t)\ \tau(t,d) \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F\text{-}measure = \frac{2 * precision * recall}{precision + recall} \quad (4)$$

$$Threshold(DP) = \min_{p \in DI} \left(\sum_{(t,w) \in \theta(p)} support\ (t) \right) \quad (5)$$

5. RESULTS AND DISCUSSION

Table1. *Shows pattern obtained after pattern discovering method. Fig.6 Shows pattern evoluation phase .*

Table1. *Term 1,Term 2,Term 3 patterns*

Item set	Patterns
1-Term	Salinity Salt tolerance Irrigation Desertification Planting Crops Agriculture India
2-Term	Agroforestry Forestry Sustainability South East Asia Asia Resource management Asia and the Pacific Highlands
3-Term	Bangladesh Villages Famine Development policies Landowners Land Social groups Landlessness

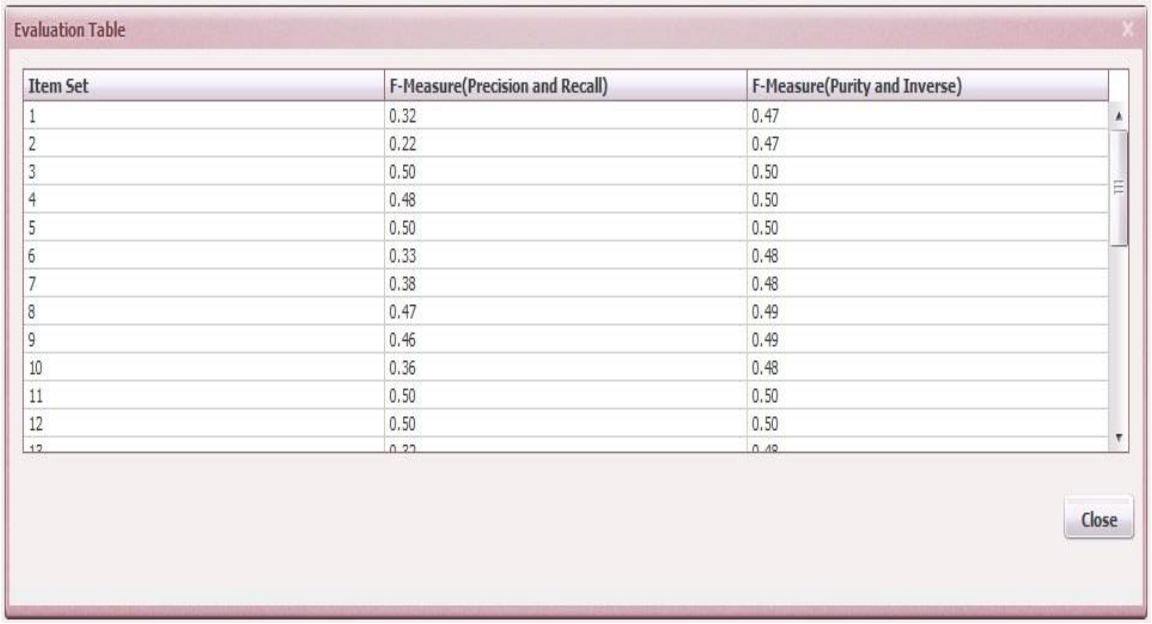

Fig2. *f-masures for each term set after pattern Evolution*

6. SYSTEM EVALUATION

After Test process, the system is evaluated using three Performance metrics precision (eq.2), recall (eq.3) and F-measure (eq.4).Using these metrics, different methods are compared to check the most appropriate method which gives maximum relevant documents to topic. Maximum number of documents relevant to topic ship are obtained at k=20. To evaluate performance of system, performance of different methods is compared using precision, recall and f1-measure. Comparison of precision, recall and f1-measure for topic ship by considering top-k documents with highest relevance score is as shown in figure 2 .Comparison of precision and recall for methods Pattern discovery, Pattern deploy and Pattern Evolving shown in figure 2 and figure 3.

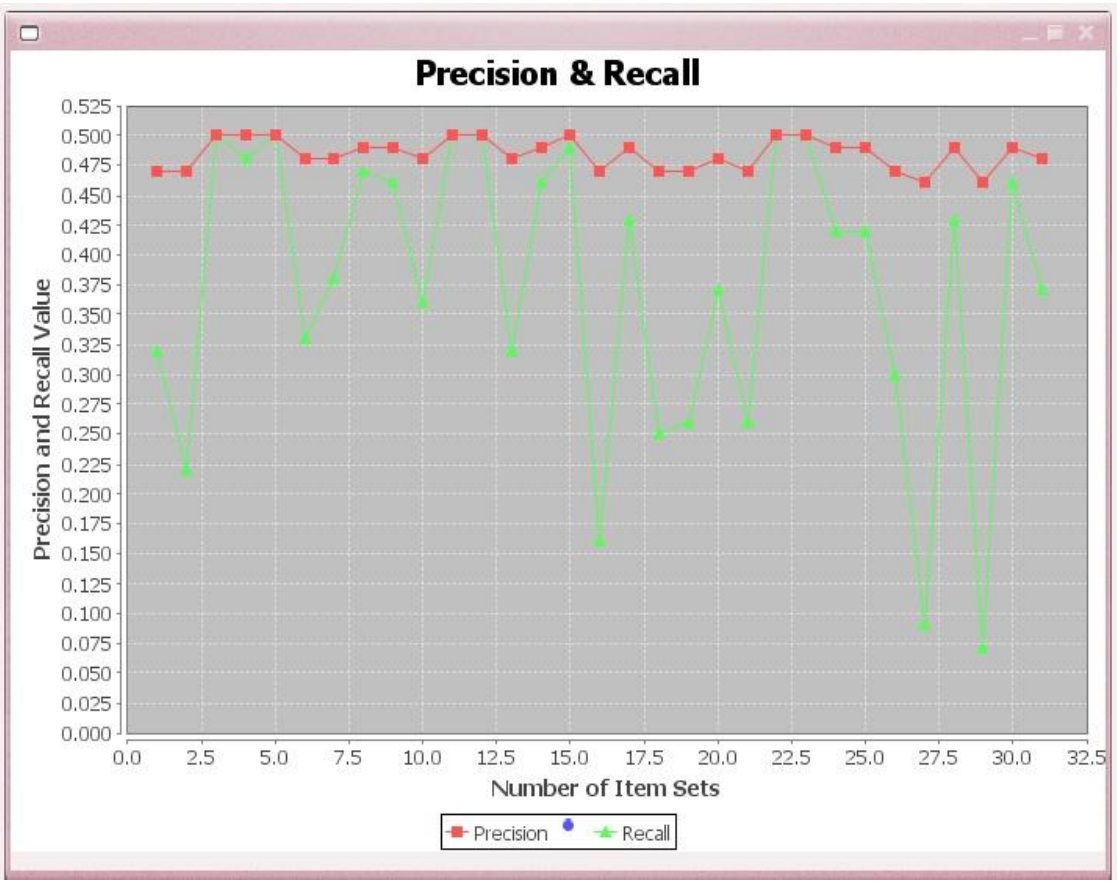

Fig3. *Precision and recall for Number of item sets*

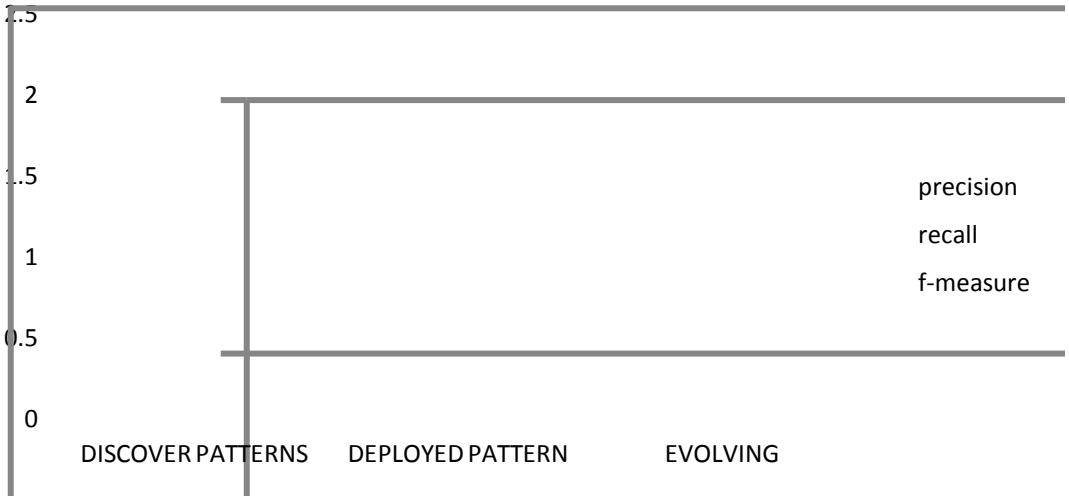

Fig4. *SCPM, PDM and DPE for RCV1- datasets*

DPE gives better results than sequential closed pattern mining (SCPM) method. So, it can be concluded that DPE and PDM are superior to SCPM.

7. CONCLUSIONS

Many text mining methods have been proposed; main drawback of these methods is terms with higher tf*idf, which are not much useful for this topic. Many data mining methods have been proposed for fulfilling various knowledge discovery tasks. These methods include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining and closed pattern mining. All frequent patterns are not useful. Hence, use of these patterns derived from data mining methods leads to ineffective performance. The reason is that some useful long patterns with high specificity lack in support (i.e., the low-frequency problem). We argue that not all frequent short patterns are useful. Hence, misinterpretations of patterns derived from data mining

techniques lead to the ineffective performance. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. An effective knowledge discovery system is implemented using three main steps: (1) discovering useful patterns by sequential closed pattern mining algorithm and non-sequential closed pattern mining algorithm. (2) Using discovered patterns by pattern deploying using PDS and PDM. (3) Adjusting user profiles by applying pattern evolution using DPE. Three performance metrics which are precision, recall and f1-measures are used to evaluate performance of the system. The result shows that an implemented system using pattern deploys and pattern evolving is superior to SCPM data mining-based method.

REFERENCES

[1] L. P. Jing, H. K. Huang, and H. B. Shi. "Improved feature selection approach tf*idf in text mining." InternationalConference on Machine Learning and Cybernetics, 2002.

[2] H. Ahonen-Myka. Discovery of frequent word sequences in text. In Proceedings of Pattern Detection and Discovery,pages 180–189, 2002.34, 61.

[3] E. Brill and P. Resnik. "A rule-based approach to prepositional phrase attachment disambiguation". In Proceedings of the 15th International Conference on Computational Linguistics (COLING), pages 1198–1204, 1994. 34.

[4] M. F. Caropreso, S. Matwin, and F. Sebastiani. Statistical phrases in automated text categorization. Technical report, Instituto di Elaborazione dell'Informazione, Technical Report IEI-B4-07-2000, 2000.

[5] M. F. Caropreso, S. Matwin, and F. Sebastiani. Statistical phrases in automated text categorization. Technical report, Instituto di Elaborazione dell'Informazione, Technical Report IEI-B4-07-2000, 2000.

[6] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu,"Effective Pattern Discovery for Text Mining", IEEE Transaction on Knowledge and data engineering, Vol. 24, no. 1, January 2012.

[7] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. In ICML'97, pages 143–151, 1997.

[8] Kjersti Aas and Line Eikvil,"Text Categorization -a servey".june 1999. (Norwedian computingcenter).

[9] Ricardo Baeza-Yates and Berthier Ribeiro-Nero,Belo Horizonte,Brazil.-"Modern Information Retrieval" January, 1999.

[10] X. Li and B. Liu. Learning to classify texts using positive and unlabeled data. In IJCAI'03, pages 587–594, 2003.

[11] Helena Ahonen and Oskari Heinonen "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections"Published in the Proceedings of ADL'98, April 22-24, 1998 in Santa Barbara, California,USA.

[12] Susan T. Dumais "Improving the Retrieval of Information from External Sources".

[13] Huma Lodhi , Craig Saunders, John Shawe-Taylor, Nello Cristianini and Chris Watkins "Text Classification using String Kernels" Journal of Machine Learning Research 2 (2002)419-444, Submitted 12/00;Published 2/02.

[14] Nitin Jindal and Bing Liu "Identifying Comparative Sentences in Text Documents".

[15] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1–47, 2002.

AUTHORS' BIOGRAPHY

Kartik K.Nikam completed her B.E. from Visvesvaraya Technological University, Belgaum in 2011. Currently he is a post graduate fellow at Computer Science and Engineering department in Shivaji university Kolhapur.His current research interests include data mining approaches.

Mrs. Anis Fatima N. Mulla completed her Master of Engineering from Shivaji University, Kolhapur. She is working as Assistant Professor in the department of Computer Science And Engineering, ADCET,Ashta. Her current research interests include Theory of Computer Science, Image Processing and Neural Networks.