USING DATA MINING METHODS KNOWLEDGE DISCOVERY FOR TEXT MINING

D.M.Kulkarni[1], S.K.Shirgave[2]

[1, 2]*IT Department Dkte's TEI Ichalkaranji (Maharashtra), India*

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 and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. Proposed work presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information.

Keywords:-*Text mining, text classification, pattern mining, pattern evolving, information filtering.*

----------***----------

1. INTRODUCTION

Knowledge discovery is a process of nontrivial extraction of information from large databases, information that is unknown and useful for user. Data mining is the first and essential step in the process of knowledge discovery. Various data mining methods are available such as association rule mining, sequential pattern mining, closed pattern mining and frequent item set mining to perform different knowledge discovery tasks. Effective use of discovered patterns is a research issue. Proposed system is implemented using different data mining methods for knowledge discovery.

Text mining is a method of retrieving useful information from a 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. New method, 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. LITERATURE SURVEY

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 of 30 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

3. PROPOSED SYSTEM

Proposed system highlights on a software upgrade-based approach to 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 transform
- Pattern discovery
- Pattern deploy
- Pattern Evolving
- Evaluation

3.1 Data Transform

Data transform is preprocessing of document. It consists of removal of irrelevant data from documents.

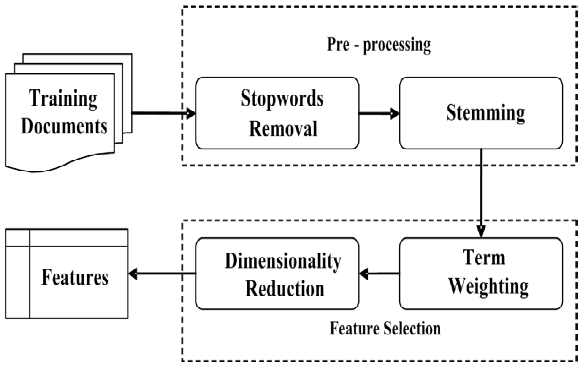

Fig 1: Data Transform

Data transform module consists of following steps as shown in figure 1.

- **Remove Stop Words**

In this step non informative words removed from document,

- **Stemming**

Stemming process to reduce derived word to its root form using Porter algorithm

- **Feature Selection**

This step assigns value to each term using a weighting scheme and removes low frequency terms.

3.2 Pattern Discovery

This module discovers patterns from preprocessed documents. Sequential closed frequent patterns as well as non sequential closed patterns are extracted using algorithms Sequential closed pattern mining and non-sequential closed pattern mining.

3.3 Pattern Deploy

Processing of discovered patterns is carried in this module. 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.

3.4 Pattern Evolving

This module removed the non meaningful patterns using deploy pattern Evolving (DPE) and Individual Pattern Evolving (IPE) Algorithms. This module finds patterns from negative document. This module identifies and removes ambiguous patterns i.e. patterns which are present in positive as well as negative documents.

3.5 Evaluation of Pattern Generated after Evolving Method

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 f1-measures.

4. EXPERIMENTAL DATASET

Several standard benchmark datasets such as Reuter's corpora, OHSUMED[5] and 20 Newsgroups [6] collection are available for experimental purposes. The most frequently used one is the Reuters dataset. Several versions of Reuter's corpora have been released. Reuters-21578 dataset is considered for experiment because it contains a reasonable number of documents with relevance judgment both in the training and test examples. Table 1 shows summary of Reuters data collections

Table 1: Summary of Reuter's data collections

Version	#docs	#trainings	#tests	#topics	Release year
Reuters-22173	22173	14,704	6,746	135	1993
Retuers-21578	21578	9,603	3,299	90	1996
RCV1	806,791	5,127	37,556	100	2000

Retuers-21578 includes 21,578 documents and 90 topics and released in 1996. Documents from data set are formatted using a structured XML scheme.

5. IMPLEMENTATION

System starts from one of the RCV1 topics and retrieves the related information with regard to the training set. Each document is preprocessed with word stemming and stops words removal and transformed into a set of transactions based on its nature of document structure. System selects one of the pattern discovery algorithms to extract patterns. Discovered patterns are deployed using one of the deploying methods, and then pattern evolving process is used to refine patterns. A concept representing the context of the topic is eventually generated. Each document in the test set is assessed by the Test module and the relevant documents to topic are shown as an output. The result of data transform is a set of transactions and each transaction consists of a vector of stemmed terms. The next step is to find frequent patterns using pattern discovery algorithms. Data mining approaches including association rule mining, frequent sequential pattern mining, closed pattern mining, and item set mining are adopted and applied to the text mining tasks. By splitting each document into several transactions (i.e., paragraphs), these mining methods are used to find frequent patterns from the textual documents. Two pattern discovery methods which have been implemented in the experiments are briefed as follows:

- **SCPM:** Finding sequential closed patterns using the algorithm SPMining. (Figure.2)

-NSPM: Finding non-sequential patterns using the algorithm.

Input:-A list of positive documents D+, minimum support (min_sup)
Output :- a set of document vectors Δ

1) $\Delta=\Phi$
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) $\vec{d} = \Phi$
6) **For each** pattern p in SP **do begin**
7) $\vec{d} = \vec{d} \oplus P'$ // p' is expanded form of p
8) **End for**
9) $\Delta = \Delta \cup \{\vec{d}\}$
10) **End for**

Fig 2: Algorithm for Sequential closed Pattern mining

PDM uses sequential or non sequential closed pattern and gives document vectors as output.PDS used sequential or non sequential closed patterns and gives document vectors with support as output.

Input:-1term patterns sequential pattern PL, minimum support (min_sup).
Output:- a set of frequent sequential patterns SP.

1) SP=Φ
2) **For each** pattern in P in PL **do begin**
3) Create set of sequences which is made of postfixes of P as PD
4) **For each** t in PD **do begin**
5) P' = P t
6) If (sup (p')>=min_sup)
7) SP=SP **U P'**
8) End if
9) If (sup(SP)<=sup(P))
10) P is closed pattern
11) End for
12) End for

Fig 3: Algorithm for Pattern deploy Method

The PDM is used with the attempt to address the problem caused by the inappropriate evaluation of patterns, discovered using data mining methods. Data mining methods, such as SPM and NSPM, utilize discovered patterns directly without any modification and thus encounter the problem of lacking frequency on specific patterns. Processing of discovered patterns is carried in this module. These discovered patterns are organized in a specific format. There are two choices for pattern deploying. One is using pattern deploying method (PDM Figure 3) and other pattern deploying with support algorithms. PDM organizes discovered patterns in <term, support> form by combining all discovered pattern vectors. PDS gives same output as PDM with support of each term. After patterns deploy, the concept of topic is built by merging patterns of all documents. While the concept is established, the relevance estimation of each document in the test dataset is conducted using the document evaluating function as shown eq. (1) in Test process. Documents in the dataset are ranked according to their relevance scores .After testing; system's performance is evaluated using the metrics such as precision, recall and f1-measures. . Deploy pattern Evolving (DPE algorithm Figure 4) is used by this module. It takes document vectors from PDM or PDS and removes the non meaningful patterns. Output of DPE is normalized document vectors. Here patterns from negative documents are identified and noisy (ambiguous) patterns i.e. Patterns which are present in Positive as well as negative documents, are filtered. Result of pattern evolving is patterns in <term, support> form by combining all deployed pattern vectors. The concept of topic is built by merging patterns of all documents While the concept is established, the relevance estimation of each document in the test dataset is conducted using the document evaluating function as shown eq.(1) in Test process. Documents in the dataset are ranked according to their relevance scores. After testing system's performance is

evaluated using the metrics such as precision, recall and f1-measures.

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

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

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

$$F1\text{-}measure = \frac{2*precision*recall}{precision+recall} \quad (4)$$

For checking performance of the system this module calculates precision, recall and f1-measure metrics. The precision is the fraction of retrieved documents that are relevant to the topic, and the recall is the fraction of relevant documents that have been retrieved. For a binary classification problem the judgment can be defined within a contingency table as depicted in Table 2.

Table 2: Contingency table

	human judgment		
System judgment	Yes		No
	Yes	TP	FP
	No	FN	TN

According to the definition in Table (2), the precision and recall are calculated using following equations. Where TP (True positives) is the number of documents the system correctly identifies as positives; FP (False Positives) is the number of documents the system falsely identifies as positives; FN (False Negatives) is the number of relevant documents the system fails to identify. The precision of first K returned documents top-K is calculated. The precision of top-K returned documents refers to the relative value of relevant documents in the first K returned documents.

Input: - A list of deployed patterns, a list of positive and negative documents, D+ and D−.

Output: - A set of term weight pairs $\vec{d}$

1) $\vec{d} = \Phi$

2) T = Threshold (D^+) **for each document d in D+ do begin**

3) **For each** negative document *nd* in D^- **do begin**

4) **If** Threshold ($\{nd\}$) > T **then**

5) $\Delta_p = \{dp \in \Omega \mid termset(dp) \cap nd \neq 0\}$

6) **Shuffling (nd,** Δ_p)

7) **End if**

8) **For each** deployed pattern *dp* in Ω **do begin**

9) $\vec{d} = \vec{d} \oplus dp$

10) **End for**

11) **End for**

Fig 4: Algorithm for pattern evolving method

The value of K use in the experiments is 20.Another metric F1-measure is calculated using following equation. To evaluate performance of system precision, recall and f1-measure of three processes is compared.

6. RESULTS OBTAINED

Following Table 3 shows pattern obtained after pattern discovery method for topic ship. Table 4 shows patterns obtained after pattern Evolving method for topic ship.

Table 3:-1-term, 2-term, 3-term patterns

	Patterns
1-term	pct offer river ship strike seamen sector redund
2-term	offer pct offer river strike pai pai seamen strike seamen
3-term	pct river offer pai strike seamen ship sourc capac industri ship japan

Table 4:-Patterns after pattern Evolving

Document no	Term	Support
23	River	1.0
43	Ship	0.25
54	Seamen	0.25
62	Missil	0.25
63	Yard	1.0
81	Industry	0.25
62	Sourc	0.25
98	Shell	1.0
98	Strike	1.0
128	Protect	1.0

7. SYSTEM EVALUATION

After Test process, the system is evaluated using three performance metrics precision (eq.2), recall (eq.3) and F1-measure (eq.4).Using these metrics, different methods are compared to check the most appropriate method which gives maximum relevant documents to topic. Reuters-21578 dataset consist of 90 topics. Comparison of precision, recall and f1-measure for topic ship by considering top-k documents with highest relevance score is as shown in figure 5. It can be observed that if value of k in top-k is chosen as 20 then system gives maximum values for precision, recall and f1-measure.

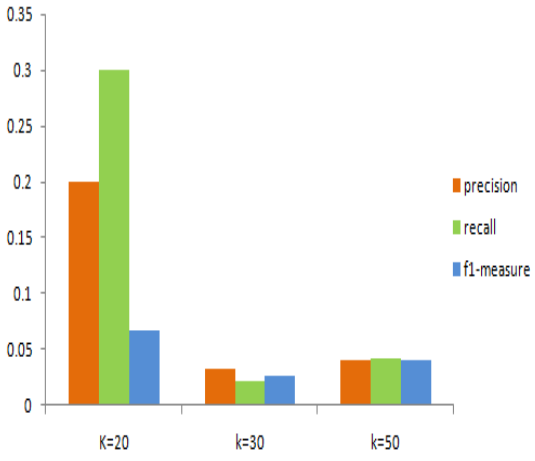

Fig 5:-Precision, recall, f1-measure for topic ship

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 and recall for methods Pattern discovery, Pattern deploy and Pattern Evolving (for topic ship is as shown in figure 6.

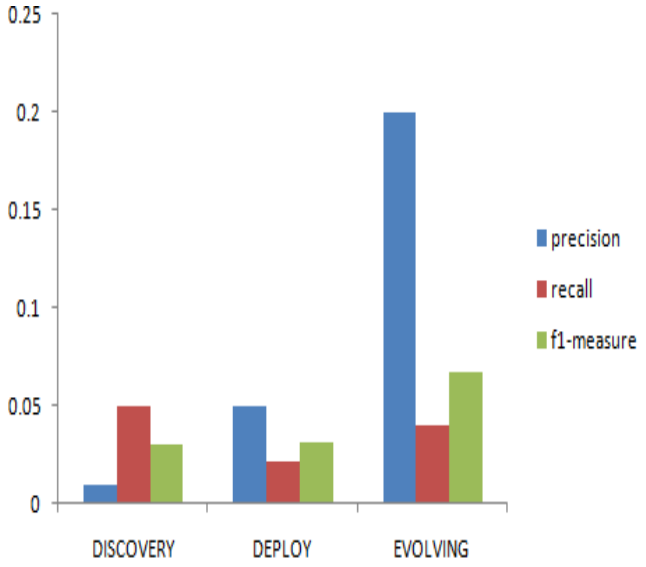

Fig 6:-SCPM, PDM and DPE for topic ship

It can be observed that maximum values for precision, recall and f1-measure are obtained from DPE. DPE gives maximum number of documents from test set that are relevant to topic ship. DPE gives better results than sequential closed pattern mining (SCPM) method. So, it can be concluded that DPE and PDM are superior to SCPM.

CONCLUSIONS

Many text mining methods have been proposed; main drawback of these methods is terms with higher tf*idf are not useful for finding concept of 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. Knowledge discovery with PDM and DPE have been proposed to overcome the above mentioned drawbacks. 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. Numerous experiments within an information filtering domain are conducted. Reuters-21578 dataset is used by the system. Three performance metrics precision, recall and f1-measures are used to evaluate performance of system. The results show that the implemented system using pattern deploy 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.” International Conference 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]. H. Ahonen, O. Heinonen, M Klemettinen, and A. I. Verkamo. “Mining in the phrasal frontier”. In Proceedings of PKDD, pages 343–350, 1997. 34, 39, 62
[5]. W. Hersh, C. Buckley, T. Leone, and D. Hickman. “Ohsumed: an interactive retrieval evaluation and new large text collection for research”. In Proceedings of the 17th ACM International Conference on Research and Development in Information Retrieval, pages 192–201, 1994.
[6]. K. Lang. News weeder: Learning to filter net news. In Proceedings of ICML, pages 331–339, 1995.