Pattern Deploy Method for Knowledge Extraction from Text Document

D.M.Kulkarni
IT Department Dkte's TEI Ichalkaranji (Maharashtra), India

Prof.S.R.Patil
IT Department Dkte's TEI Ichalkaranji (Maharashtra), India

Prof.T.I.Bagban
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— Terms—Text mining, text classification, pattern mining, pattern evolving, information filtering.

I. 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.

II. 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.

III. 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

Data transform

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

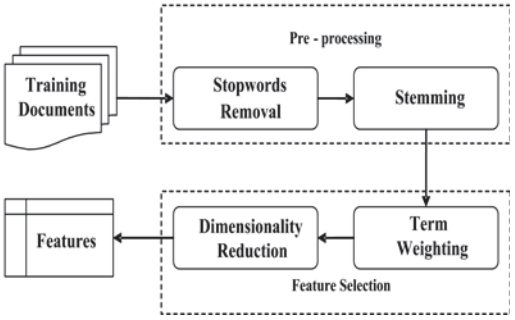

Figure 1.1: Data Transform

Data transform module consists of following steps as shown in figure 1.1

- 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.

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.

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.

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.

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.

IV. 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 1.1** shows summary of Reuters data collections

Table 1.1: Summary of Reuters 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.

V. SYSTEM EVALUATION

After Test process, the system is evaluated using three performance metrics precision recall and F1-measure.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 **figure1. 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.

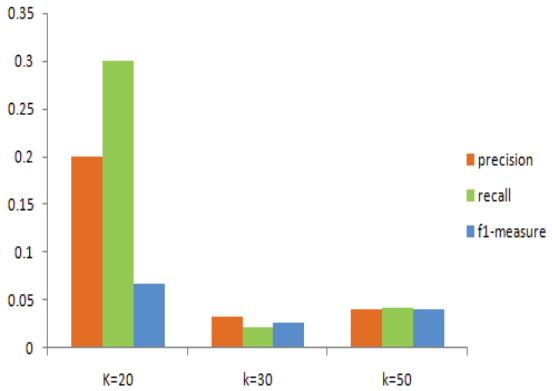

Figure 1.5:-Precision, recall, f1-measure for topic ship

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 1.6.**

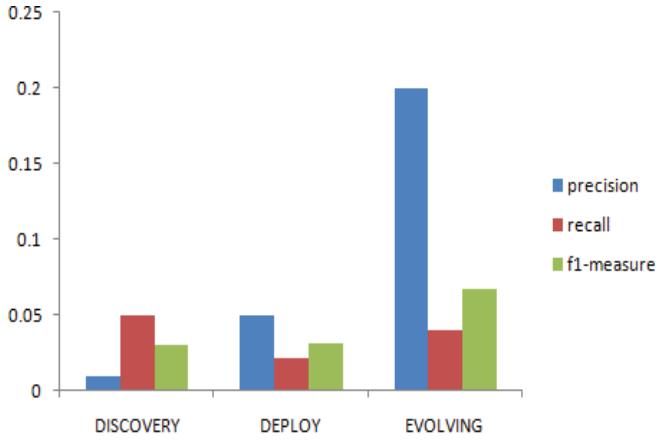

Figure 1.6:-SCPM, PDM and DPE for topic ship

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.

VI. CONCLUSION

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.*