International Journal of Research (IJR) Vol-1, Issue-11 December 2014 **ISSN 2348-6848**

# Analysis of Effective Pattern Discovery with Text Mining in Business Based Application

Ms. Rupali Bhaisare [1] & Prof. Vinod Nayyar [2]

[1]Nagpur University, Department of M. Tech Computer Science and Engineering, Nagpur, Maharashtra, India
*rupali.bhaisare786@gmail.com*

[2] Nagpur University, Department of M. Tech Computer Science and Engineering, Nagpur, Maharashtra, India
*vinodnayyar5@gmail.com*

## Abstract:

*In text documents data mining techniques have been proposed for mining useful patterns. But how to effectively use and update discovered patterns is still an open research issue, especially in the text mining. So existing text mining methods adopted term-based approaches but they all suffer from the problems of polysemy and synonymy. This paper presents an effective pattern discovery technique, which first calculates discovered specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns rather than the distribution in documents for solving the misinterpretation problem. It also considers the influence of patterns from the negative document examples to find noisy patterns and try to reduce their influence for the low-frequency problem. This approach can improve the accuracy of evaluating term weights because discovered patterns are more specific than whole documents.*

## Keywords:

Text mining; Text classification; Information filtering; Clustering

### 1. Introduction

From the past years, a number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include sequential pattern mining, association rule mining, maximum pattern mining, frequent item set mining, and closed pattern mining. There is rapidly growth of digital data is available in recent years; knowledge discovery and data mining have work together this is a great deal of attention for need of meaningful data into useful information and knowledge. Data mining is efficient step in the process of knowledge discovery in databases. And Text mining is used to finding relevant & interesting information from huge database. Text mining is to exploit information contained in extual documents in various ways and including with discovery of patterns, association among entities, etc.

### 1.1 Text mining

Text analytics sometimes called as text mining is one way to make unstructured data usable by a computer. Text Mining is the discovery by computer of new, previously unknown information, by automatically extracting and relating information from different written resources. Text mining is the discovery of interesting knowledge in text documents. Text mining is to exploit information contained in textual documents in various ways including discovery of patterns, association among entities, predictive rules etc. And we can say text mining is used to finding relevant & interesting information from huge amount of database.

### 1.2 Text classification

Text classification is the process of classifying documents into predefined categories based on their content. It is the automated assignment of natural language texts to predefined categories. It also can be viewed as the process of finding a proper method to distinguish data classes or concepts. Text classification is the primary requirement of text retrieval systems, which retrieve texts in response to a user query, and text understanding systems, which transform text in some way such as producing summaries, answering questions or extracting data.

### 1.3 Information filtering

Information filtering system removes unwanted information from an information stream using automated or computerized methods to user. Its main goal is the management of the information overload and increment of the semantic signal-to-noise ratio. Information filtering is usually works by specifying character strings, if they matched, then it indicate undesirable content that is to be screened out.

### 1.4 Clustering

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 . In other words, clustering aims for maximizing the intra-class similarity and minimizing the inter-class similarity.

## 2. Problem Statement

A. The term-based approach is suffered from the problems of polysemy and synonymy.
B. Low frequency problem of words.

## 3. Project Objective

The objectives are showing how the proposed approach can help improving the effectiveness of pattern-based approaches:

- Pattern deploying method has better performance for the interpretation of discovered patterns in text documents. This deploying approach is not only promising for pattern-based approaches but also significant for the concept-based model.
- Pattern evolution is designed to achieve the high performance for determining relevant information to answer what users want. The method would be better than other pattern-based, concept-based and term-based models in the effectiveness.

## 4. Proposed System Overview

Proposed system works on text clustering has a frequent concept to cluster the text documents. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. Our Proposed algorithm utilizes the semantic relationship between words to create concepts. Associating a meaningful label to each final cluster is more essential. Then, the high dimensionality of text documents should be reduced. Only authorized user retrieve the particular dataset for pre-processing and apply three methods. When complete the process with three methods it will provide data with term set and their experimental coefficient to user as per their selected dataset.

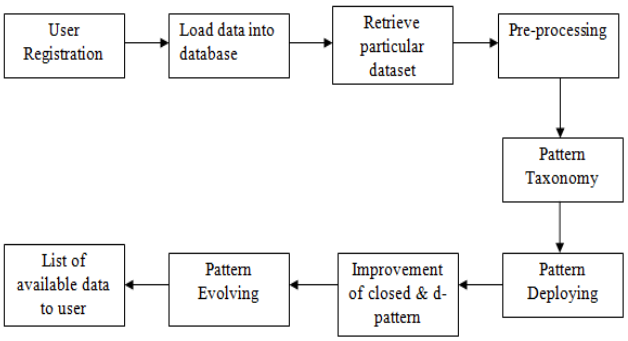


**Figure 1:** System Architecture of Proposed Work.

## 5. Methodology

### 5.1 Pattern Taxonomy Method (PTM)

Two main stages are considered in PTM. The first stage is how to extract useful phrases from text documents and second stage is how to use

these discovered patterns to improve the effectiveness of a knowledge discovery system. In PTM, we split a text document into a set of paragraphs and treat each paragraph as an individual transaction, which consists of a set of word terms. At the subsequent phase, we apply the data mining method to find frequent patterns from these transactions and generate patterns taxonomies.

**Table 1:** Example of PTM

| Paragraph | Terms |
|---|---|
| dp1 | t4t5 |
| dp2 | t5t6t7 |
| dp3 | t2t3t4t5 |

### 5.2 Pattern Deploying Method (PDM)

In the pattern taxonomy semantic information is help to improve the performance of closed patterns in text mining, and then we need to interpret discovered patterns by summarizing them as d-patterns in order to accurately evaluate term weights. We started from the documents in the case of textual data, several pattern taxonomies can be built by finding informative patterns using data mining methods. On the other hand, a feature space which consists of a set of individual terms is generated by the use of traditional document indexing techniques. Then the next step, created pattern taxonomies and feature space can be used to represent the concept of documents by applying a data mining-based method like SPM. By deploying patterns into the feature space, PDM use of sequential patterns to keep the useful semantic information but also improve the system efficiency by preventing the time-consuming pattern discovery approaches from being used in the phrase of document evaluation.

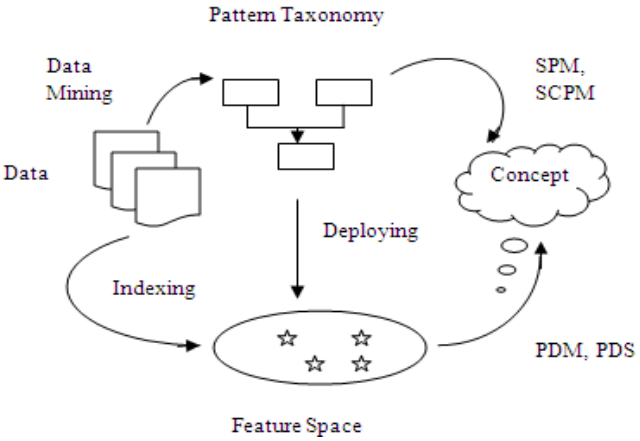


**Figure 2:** Flowchart of pattern deploying methods in pattern taxonomy method.

**Table 2:** Example of PDM

| Frequent Pattern | Covering Set |
|---|---|
| {t2,t4,t6} | {dp1,dp2,dp3} |
| {t1,t4} | {dp2,dp3,dp4} |
| {t4} | {dp2,dp3,dp4} |
| {t1,t2} | {dp1,dp5,dp6} |
| {t6} | {dp2,dp3,dp4,dp5,dp6} |

### 5.3 Pattern Evolution Method (PEM)

5.3.1. Deployed Pattern Evolution (DPE)

The PTM has been significantly improved after the adoption of pattern deploying method, which uses the strategy of mapping discovered patterns into a hypothesis space for solving the low-frequency problem to the specific long patterns. There is sometime negative documents contain some useful information to identify ambiguous patterns in the concept.

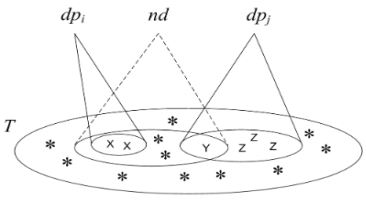


**Figure 3:** A negative document nd and its offending deployed patterns.

A negative document nd is a document that the system sometime identified as a positive document. The offender of nd is a deployed

pattern which obtains at least one component that appears in nd. The set of offenders of nd is defined by:

$$\Delta_p - \{ dp \epsilon \Omega \mid termset(dp) \cap nd \neq \theta \}$$

Figure 3 illustrates the relationship between a negative document nd and its offenders.

5.3.2 Individual Pattern Evolution (IPE)

In 5.3.1, a pattern refinement strategy is proposed using the pattern evolving approach DPE to reshuffle the weight distribution within offenders. Essentially, it is reasonable that the pattern evolution is applied to a pattern which appears both in the offender and the negative document for the purpose of removing the suspicious source of "noise". Therefore, an evolving approach called Individual Pattern Evolution (IPE) is proposed in this section.

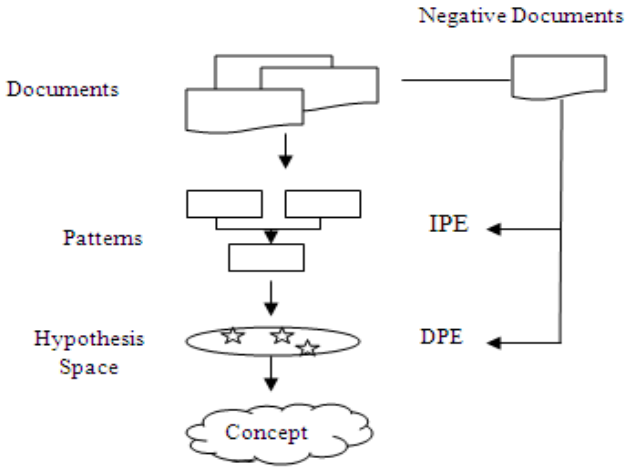


**Figure 4:** The flowchart of two pattern evolving approaches.

Figure 4 illustrates the different states in which the evolution of patterns takes place using DPE and IPE. When a negative document is detected, DPE starts to find offenders and implements pattern evolving at "Hypothesis Space" state. In contrast, IPE executes the same action at "Pattern" state. In addition, the structures of "Hypothesis Space" and "Pattern" are different, and thus an alternative definition and algorithm for IPE are needed.

## 6. Result Analysis

This section describes the experimental evaluation of proposed approaches featured in the pattern taxonomy model. Most of the standard performance measures (i.e. precision, recall) are used for evaluating the experimental performance. The PTM model comprises the methods including pattern discovery approaches, pattern deploying methods, and pattern evolution strategies. The experimental results are compared with other well-know IF-related methods including, Probabilistic method (Prob) and Rocchio method.

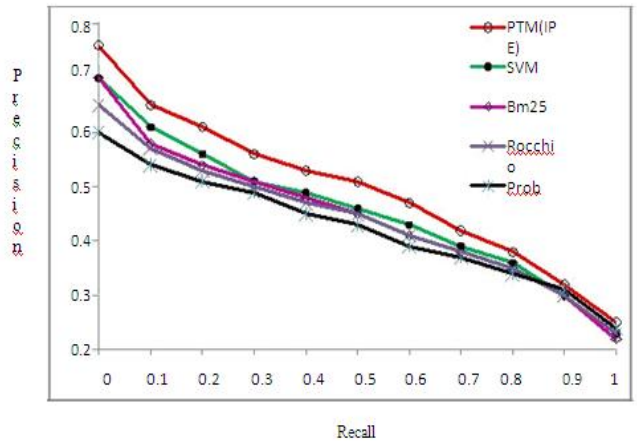


**Figure 5:** Comparing PTM (IPE) with term-based methods on the first 50 RCV topics.

Fig. 5 presents the plotting of precisions standard points for PTM and several term-based methods on the first 50 RCV topics. The difference of performance for all methods is easier to be recognized in this figure. Again, the PTM method outperforms all other methods, including Rocchio, Prob, BM25 and SVM methods. Among these methods, the Prob method achieves a noticeable score of precision at the 0.6 point where recall equals to zero, meaning that the Prob method is able to promote top relevant documents. The overall performance of PTM (IPE) is better than that for Prob method.

## 7. Conclusions

This paper presents the research on the concept of developing an effective knowledge discovery model (PTM) based on pattern taxonomies. PTM is implemented by three main steps: (1) discovering useful patterns by integrating

sequential closed pattern mining algorithm (2) using discovered patterns by pattern deploying; (3) adjusting user profiles by applying pattern evolution. Various mechanisms in each step are proposed and evaluated for fulfilling the PTM model. The latest version of the Reuters dataset, RCV, is selected and tested by the proposed PTM-based information filtering system. The results show that the PTM outperforms not only several pure data mining-based methods, but also traditional probabilistic and Rocchio methods.

## Acknowledgments

I would like to thanks my guide Prof. Vinod Nayyar Sir to giving me their excellent knowledge, insightful comments and suggestions for making this paper.

## References

[1] Ning Zhong, Yuefeng Li and Sheng Tang Wu,"Effective Patterns Discovery for Text Mining," IEEE Transaction On Knowledge And Data Engineering, Vol. 24, No.1, January 2012.

[2] Y. Li, S-T. Wu, and Y. Xu, Deploying Association Rules on Hypothesis Spaces, Proceeding of International Conference on Computational Intelligence for Modelling Control and Automation, 2004.

[3] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules in Large Databases," Proc. 20th Int'l Conf. Very Large Data Bases (VLDB '94), pp. 478-499, 1994

[4] M. Goebel and L. Gruenwald. A survey of data mining and knowledge discovery software tools. SIGKDD Explorations, 2003.

[5] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R.Y. Lau, "A Two-Stage Text Mining Model for Information Filtering," Proc. ACM 17th Conf. Information and Knowledge Management, 2008.

[6] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management: An Int'l J., vol. 24, no. 5, 1988.

[7] H. Ahonen, O. Heinonen, M.Klemettinen, and A.I. Verkamo. Mining in the phrasal frontier. In Proceedings of PKDD, 1997.

[8] S-T. Wu, Y. Li, and Y Xu. Deploying approaches for pattern refinement in text mining. In Proceedings of ICDM , 2006.

## Author Profile

**Rupali Bhaisare** has received the B. E. degree in Computer Technology from RTMNU, Maharashtra, India in 2010 & pursuing M. Tech in CSE from RTMNU. Since, 2011 she is working in the department of IT as a Asst. Professor in SRMCEW, Nagpur, Maharashtra, India.