ISSN: 2348 9510

**International Journal Of Core Engineering & Management (IJCEM)**
**Volume 2, Issue 10, January 2016**

# D-PATTERN DISCOVERY AND TERM SUPPORT EVALUATION FOR EFFECTIVE TEXT MINING

**Ms. Awantika Bijwe**

Savitribai Phule Pune University, Pune, Maharashtra, India

*awantika.bijwe@gmail.com*

## *ABSTRACT*

***Data Mining (often known as knowledge discovery) is the process of identifying patterns in large sets of data. Useful knowledge may be hidden in the text documents. If we extract this knowledge it may provide good support for planners, decision makers, and legal institutions or organizations. In text documents various data mining techniques have been proposed for mining useful patterns. But there are some questions, how to effectively use and update discovered patterns is still a research issue, especially in the text mining. Most of the existing text mining methods adopted term-based data mining approaches but they all suffer from the problems of polysemy and synonymy. Our proposed system implements 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 with effective patterns as per the users requirements.***

***Keywords: Data mining, Effective pattern discovery, pattern evolving, Pattern deploying***

## 1. INTRODUCTION

ISSN: 2348 9510

**International Journal Of Core Engineering & Management (IJCEM)**
**Volume 2, Issue 10, January 2016**

Data Mining is the process of discovering new correlations, patterns, and trends by mining large amounts of data stored in warehouses, using artificial intelligence, statistical and mathematical techniques. Data mining is the principle of sorting through large amounts of data and picking out required information. It has been described as finding hidden information in a database. Alternatively, it has been called exploratory data analysis, data driven discovery, and deductive learning and the science of extracting useful information from large databases. Data mining is usually used by business intelligence organizations, and financial analysts, but it is increasingly used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods. One of major area where data mining can be used in the industry is in monitoring systems.

Text mining is to exploit information present in textual documents in various ways including discovery of patterns, association among entities, etc. Earlier, Information Retrieval (IR) provided many term-based methods to solve this challenge, The merits of term-based methods include efficient computational performance as well as some theories for term weighting, which have emerged over the last couple of decades from the IR and machine learning communities. But the problems with term based methods are polysemy and synonymy, where polysemy means a particular word has multiple meanings, and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users really want. After that, people have often held the hypothesis that phrase-based approaches may perform better than the term based ones, as phrases may carry more "semantics" like information. This hypothesis has not fared too well in the history of Data mining. Although phrases are less ambiguous and more discriminative than individual terms, the likely reasons for the discouraging performance include:

1) Phrases have inferior statistical properties to terms

ISSN: 2348 9510

**International Journal Of Core Engineering & Management (IJCEM)**
**Volume 2, Issue 10, January 2016**

2)They have very low frequency of occurrence

3) There are large numbers of redundant and noisy phrases might present in it.

In the presence of these setbacks, sequential patterns used in data mining community have turned out to be a promising alternative to phrases because sequential patterns enjoy good statistical properties like terms. To overcome the disadvantages of phrase-based approaches, pattern mining-based approaches have been proposed, which adopted the concept of closed sequential patterns, and pruned non closed patterns.

## 2. BACKGROUND STUDY

**Data Mining**

Data mining is the process of finding patterns among dozens of fields in large relational databases. Data mining is primarily used today by companies with a strong consumer focus on financial, retail, communication, and marketing organizations. It enables these companies to determine relationships among internal factors such as product positioning, price, or staff skills, and external factors such as economic indicators, competition, and customer demographics. It helps to company to determine the impact on sales, customer satisfaction, and profits. Data mining software analyzes relationships and patterns in stored transaction database.

**Sequential patterns**:

In sequential patterns data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

**Clustering**

ISSN: 2348 9510

**International Journal Of Core Engineering & Management (IJCEM)**
**Volume 2, Issue 10, January 2016**

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar properties to each other than to those in other groups . It's a main task of retrieving data from data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, image analysis, and pattern recognition information retrieval. Work with clustering helps to modify data preprocessing and model parameters until the result achieves the required properties.

**Information Filtering**

Information filtering is a process that removes redundant or unwanted information from an information stream using automated or computerized methods to user. Its main aim is the management of the information overload and increment of the semantic signal-to-noise ratio. In this the user's profile is compared to some reference characteristics. Information filtering is usually works by specifying character strings, if they matched, then it indicate undesirable content that is to be filtered out.

**Association Rules**

In data mining, association rule learning is a popular method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Association rules are if/then statements that help to uncover relationships between seemingly unrelated data in a relational database. Association rules are created by analyzing data for frequent if/then patterns and using the criteria support and confidence to identify the most important relationships.

**Pattern Mining**

ISSN: 2348 9510

**International Journal Of Core Engineering & Management (IJCEM)**
**Volume 2, Issue 10, January 2016**

Patterns are item subsequences, sets, or substructures that appear in a data set with frequency no less than a user specified threshold. For example, a set of items, such as milk and bread that appear frequently together in a transaction data set is a frequent item set. A subsequence, such as buying first a PC, then a digital camera, and then a memory card, if it occurs frequently in a shopping history database, is a frequent pattern .A substructure can refer to different structural forms, such as sub graphs, sub trees, or sub lattices, which may be combined with item sets or subsequence.

## 3. LITERATURE REVIEW

1. Fast Algorithms for Mining Association Rules in large databases [1]

In this paper the problem of discovering association rules between items in a large database of sales transactions. For solving this problem author also show how the best features of the two proposed algorithms can be combined into a hybrid algorithm, called AprioriHybrid algorithm.

2. Kernel methods for document filtering [2]

In this paper algorithms implemented by KERMIT IST European project is concerned with the investigation of kernel methods for applications related to the categorization, retrieval, clustering and ranking of text documents and of images. Author overcomes the problem with lack of improvement in performance when polynomial kernels of degree higher than one or radial basis function kernels.

3. Mining Generalized Association Rules [3]

In this paper author introduces the problem of mining generalized association rules with a large database of customer transactions, where each transaction consists of a set of items, and taxonomy on the items. Here solution to the problem is to replace each transaction with an “extended transaction” that contains all the items in the original transaction.

ISSN: 2348 9510

**International Journal Of Core Engineering & Management (IJCEM)**
**Volume 2, Issue 10, January 2016**

4. Learning to Classify Texts Using Positive and Unlabeled Data [7]

Currently textual documents are increasingly added to the World Wide Web and also the electronic databases of organizations. One of the representations which are well known is known as bag of words approach it makes use of keywords. Tf*idf weighting scheme is presented in for representing text.

## 4. METHODOLOGY

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.

The method also includes a pattern taxonomy model to discover patterns and pattern deploying methods to update discovered patterns based on their frequency. And an individual pattern evolution technique is used to discover the concept of each area for the classification process.

The various modules of Effective pattern discovery are as follows

**Data preprocessing**

This process involves data cleaning and noise removal. It also involves 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 Word removal:* Stop words are the words which are filtered out prior to, or after, processing of natural language data. In this step non informative words removed from document.

ISSN: 2348 9510

**International Journal Of Core Engineering & Management (IJCEM)**
**Volume 2, Issue 10, January 2016**

*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.

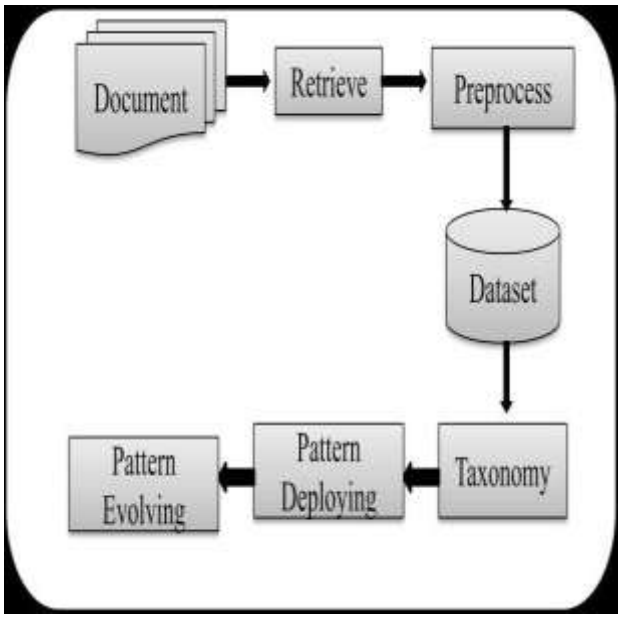



Fig.1 - System flow diagram

ISSN: 2348 9510

International Journal Of Core Engineering & Management (IJCEM)
Volume 2, Issue 10, January 2016

**Pattern taxonomy model/Pattern discovery**

In PTM, we split a text into set of paragraphs and treat each paragraph as an individual transaction, which consists of a set of words. At the subsequent phase, apply the data mining method to find frequent pattern from these transaction and generate pattern taxonomies. During the pruning phase, non-meaning and redundant pattern are eliminated by applying a proposed pruning scheme. Pattern taxonomy is a tree-like structure that illustrates the relationship between patterns extracted from a text collection.

*Closed Sequential Pattern*: A frequent sequential pattern *P* is a closed sequential pattern if there exist no frequent sequential pattern

**Pattern deploying**

The discovered patterns from first are summarized in pattern deploying module. 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.

**Pattern evolving**

In pattern 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.

ISSN: 2348 9510

**International Journal Of Core Engineering & Management (IJCEM)**
**Volume 2, Issue 10, January 2016**

**Evaluation**

This module 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.

## 5. TECHNICAL OVERVIEW

### A. D-PATTERN MINING ALGORITHM

To improve the efficiency of the pattern taxonomy mining, an algorithm, PMining, was proposed in [50] to find all closed sequential patterns, which used the well-known Apriori property in order to reduce the searching space. Algorithm 1 (PTM) shown in Figure describes the training process of finding the set of d-patterns. For every positive document, the SPMining algorithm is first called in step 4 giving rise to a set of closed sequential patterns SP. The main focus of this paper is the deploying process, which consists of the dpattern discovery and term support evaluation. In Algorithm 1 all discovered patterns in a positive document are composed into a dpattern giving rise to a set of d-patterns DP in steps 6 to 9.

**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 $\widehat{d} = \emptyset$;
6 **foreach** *pattern* $p_i \in SP$ **do**
7 $p = \{(t, 1) | t \in p_i\}$;
8 $\widehat{d} = \widehat{d} \oplus p$;
9 **end**
10 $DP = DP \cup \{\widehat{d}\}$;
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**

ISSN: 2348 9510

**International Journal Of Core Engineering & Management (IJCEM)**
**Volume 2, Issue 10, January 2016**

B. INNER PATTERN EVOLUTION

In this section, we discourse how to reorganization supports of terms within common forms of d-patterns on the bases of negative ocuments in the training set. The technique will be useful to reduce the side effects of noisy patterns because of the Fig. 3 Algorithm

ISSN: 2348 9510

**International Journal Of Core Engineering & Management (IJCEM)**
**Volume 2, Issue 10, January 2016**

IPEvolving Fig. 4. Algorithm Shuffling low-frequency problem. This technique is called inner pattern evolution here, because it only changes a patterns term supports within the pattern. A inception is usually used to classify documents are into relevant or irrelevant categories.

$$weight(d) = \sum_{t \in T} support(t)\tau(t,d),$$

where $support(t)$ is defined in Algorithm 1 (Fig. 2); and $\tau(t,d) = 1$ if $t \in d$; otherwise $\tau(t,d) = 0$.

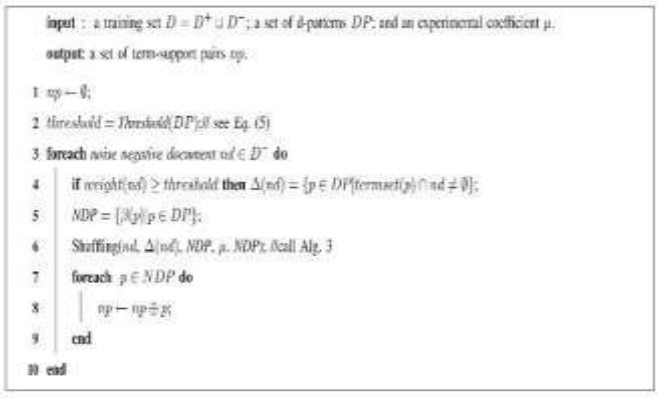
**Advantages**

1. In this system we are giving high priority for long sequence in the evaluated pattern.

2. In this system we are giving term weight based on occurrence of term in long pattern (sequence).

3. It improves the effectiveness of using and updating discovered patterns for finding relevant and interesting information

4. The proposed approach is used to improve the accuracy of evaluating term weights. Because, the discovered patterns are more specific than whole documents



ISSN: 2348 9510

**International Journal Of Core Engineering & Management (IJCEM)**
**Volume 2, Issue 10, January 2016**

## 6. CONCLUSION

Many data mining methods have been proposed so far for fulfilling various knowledge discovery tasks. These methods are frequent item set mining, association rule 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.
- So, misinterpretations of patterns derived from data mining techniques lead to the ineffective performance. The proposed system 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 major 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 (3)Adjusting user profiles by applying pattern evolution.

ISSN: 2348 9510

**International Journal Of Core Engineering & Management (IJCEM)**
**Volume 2, Issue 10, January 2016**

## REFERENCES

[1] R. Agrawal "Fast Algorithms for Mining Association Rules in Large Databases," Very Large Data Bases (VLDB '94), pp. 478-499, 1994.

[2] N.Cancedda, C.gentile "Kernel methods for documents Filtering".

[3] R. Srikant and R. Agrawal, "Mining Generalized Association Rules," Very Large Data Bases (VLDB '95), pp. 407-419,1995.

[4] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval.Addison Wesley, 1999.

[5] N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders, "Word- Sequence Kernels," J. Machine Learning Research, vol. 3, pp. 1059-1082, 2003.

[6] M.F. Caropreso, S. Matwin, and F. Sebastiani, "Statistical Phrases in Automated Text Categorization," Technical Report IEI-B4-07- 2000, Instituto di Elaborazione dell'Informazione, 2000.

[7] X. Li "Learning to Classify Texts Using Positive and Unlabeled Data," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI '03), pp. 587-594, 2003.

ISSN: 2348 9510

**International Journal Of Core Engineering & Management (IJCEM)**
**Volume 2, Issue 10, January 2016**

**Author Profile**

**AWANTIKA BIJWE,** Awantika Bijwe has completed her bachelors in computer science and she has also completed her MCA from Santa Gadgebaba University, Amravati. She has an experience of more than 3.5 years of industry and 5 years in teaching various core computer application subjects like Software Engineering, Web technology, Microsoft .Net, Software project management, Research methodology, and Organizational Behavior and Computer organization.