

Knowledge Discovery Technique for Mining Text Information

Mr. Amrut Madhukar Jadhav¹

¹ Student, Computer Engineering Department, JSPM's ICOER Wagholi Pune, Maharashtra, India

Abstract - Text mining is the process of finding significant data. Extracted knowledge is coming from unstructured textual data. So the proper technique is required to implement this strategy to get important knowledge from unstructured data. In this paper, text mining approach is presented with the help of deployment, evolution process. Preprocessing is the preliminary step applied to filter data.

Key Words: text mining, pattern deployment, pattern evolving.

1. INTRODUCTION

Now a day, many organizations or companies produces large data for storing information. Such data basically present in unstructured format. So that, it is not possible to handle large volume data which requires more time to handle [1]. We require such technique which will handle this type data and finds accurate knowledge.

Text mining comprises of various functions such as question and answering by interact with user, clustering to cluster documents, topic tracking to maintain user profile, categorization to group related information, information extraction and classification[10]. Many applications, such as business management and market analysis of products, can gain by the use of the information and patterns extracted from massive amount of data. Knowledge discovery is the process of nontrivial extraction of information from huge databases and information that is indirectly presented in the data, before unknown and possibly useful for users [11].

So that data mining field becomes an essential step in the process of knowledge discovery in large data set. With a huge number of patterns produced by using data mining approaches, how to efficiently use and update patterns is still becomes research topic [14]. Proposed system is evaluates the measures of patterns using pattern deploying process as well as finds patterns from the negative training examples using pattern Evolving process. Text mining is the technique that helps users finds useful information from a large amount of digital text data. It is therefore crucial that a good text mining model should retrieve the information that users require with relevant efficiency. Traditional Information Retrieval (IR) has the 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 adequately provide users with what they really need. Many text mining methods have been developed in order to achieve the goal of retrieving useful information for users.

Most research works in the data mining community have focused on developing efficient mining algorithms for discovering a variety of patterns from a larger data collection. However, searching for useful and interesting patterns is still an open problem [13]. In the field of text mining, data mining techniques can be used to find various text patterns, such as sequential patterns, frequent item sets, co-occurring terms and multiple grams, for building up a representation with these new types of features. Nevertheless, the first problem is how to effectively deal with the large amount of patterns generated by using the data mining methods.

Using phrases for the text representation still has doubts in increasing performance over domains of text categorization tasks, meaning that there exists no particular representation method with dominating advantage over other. Instead of the keyword-based approach which is typically used by text mining-related tasks in the past, the pattern-based model (single term or multiple terms) is employed to perform the same concept of task [12]. There are two phases that we need to consider when we use pattern-based models in text mining: one is how to discover useful patterns from digital text documents, and the other is how to utilize these mined patterns to improve the systems performance [8].

In this paper technique of pattern refining approach is used. It first calculates discovered specificity 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 training examples to find ambiguous (noisy) patterns and try to reduce their influence for the low-frequency problem. The process of updating ambiguous patterns can be referred as pattern evolution [9]. The proposed approach can improve the accuracy of evaluating term weights because discovered patterns are more specific than whole documents.

2. LITERATURE SURVEY

Text mining process is basically useful to extract knowledge from text documents. To improve the process of pattern discovery, the concepts of pattern deployment, pattern evolving and shuffling has been used [1].It

presents an innovative idea for finding patterns. To measure occurrence of terms basically the concept TFIDF (term frequency-inverse document frequency) has been used [2]. Various approaches [7] are used to extract patterns such as keyword based and phrase based [1]. Phrase based performs better than keyword based because it carries more semantic. Bag of terms of words has number of problems that contains set of terms and regarding knowledge amongst a vast set of words to increase the efficiency of system. Single words carries less semantics, so ambiguity arises. To overcome such kind of problem, phrase based (having multiple words) mechanism becomes better [4] [6] [7]. So, phrases having multiple words show less ambiguity to fetch patterns. Keyword based technique becomes inadequate as compare to phrase based technique because single word is not that much sufficient to express the knowledge [1][4]. To identify groups of words that create meaningful phrases is a better method, especially for phrases indicating important concepts in the text. Clustering provides grouping of related classes [3] so that it improves representation of text. Various methods of text mining has been used now days [3][5], such as information extraction, topic tracking, clustering, classification etc.

3. PROBLEM STATEMENT AND SCOPE

3.1 Problem Statement

The main focus of this system is to discover patterns those are more relevant from document. How to effectively use those relevant patterns is becomes challenging task.

3.2 Scope

Text mining process is basically applied on unstructured data in text format. So that, user get benefits of retrieving documents using this kind of technique. Such system only manipulates textual data.

4. PROPOSED SYSTEM

The proposed system gives a Knowledge Discovery model an attempt to effective exploit the discovered patterns in a large data collection using data mining methods. This technique increases efficiency of discovered patterns using algorithms such as pattern deploying and pattern evolving. System utilizes data which is in form of text. This collection of data contains training set of documents for implementation of whole system. This data set contains positive as well as negative documents. Positive documents are those which relevant to topic else it treats as negative. Whole system is composed of data preprocessing, pattern taxonomy model, pattern deploying process, evolving and clustering mechanisms. So that proposed system is dived into four modules that represent these processes.

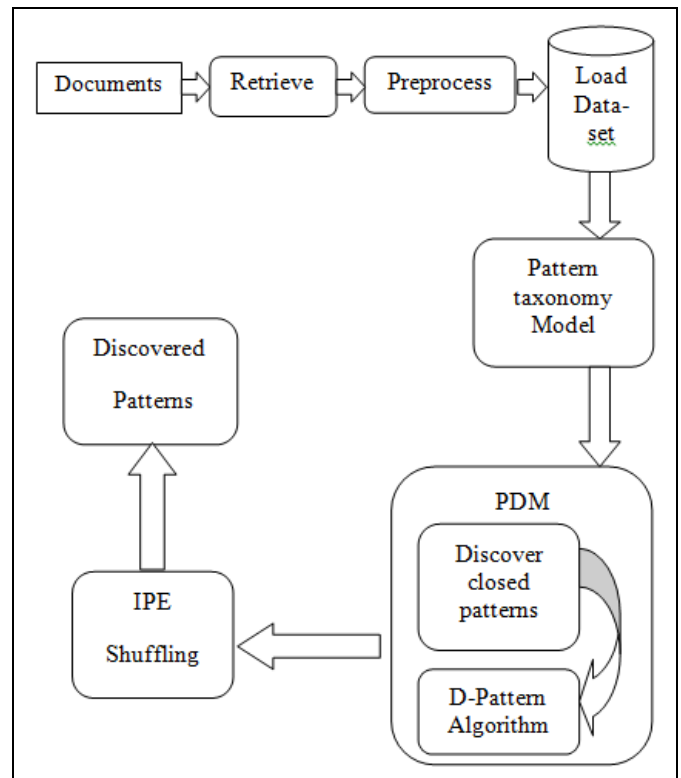


Fig -1: System Architecture

Proposed system divides the whole work into various stages v.i.z. preprocessing, pattern taxonomy model, pattern deployment and pattern evolution.

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 process, 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.

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 process, 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.

Algorithm 1: 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  $p_i \in SP$  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 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.

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 $weight(nd) \geq Threshold(DP)$. 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 .

Algorithm 2: 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  $np \leftarrow \emptyset$ ;
2  $threshold = Threshold(DP)$ ;
3 foreach noise negative documents  $nd \in D^-$  do
4     if  $weight(nd) \geq threshold$  then  $\Delta(nd) = \{p \in DP | termset(p) \cap nd \neq \emptyset\}$ ;
5      $NDP = \{\beta(p) | p \in DP\}$ ;
6     Shuffling( $nd, \Delta(nd), NDP, \mu, NDP$ );
7     foreach  $p \in NDP$  do
8          $np \leftarrow np \oplus p$ ;
9     end
10 end
    
```

Algorithm 3: 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  $\Delta(nd)$  do
2   if  $termset(p) \subseteq nd$  then  $NDP = NDP - \{p\}$ ; //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

```

5. MATHEMATICAL MODEL

5.1 Deterministic Finite Automata (DFA)

It is a state machine which accepts or rejects strings as input and produces unique

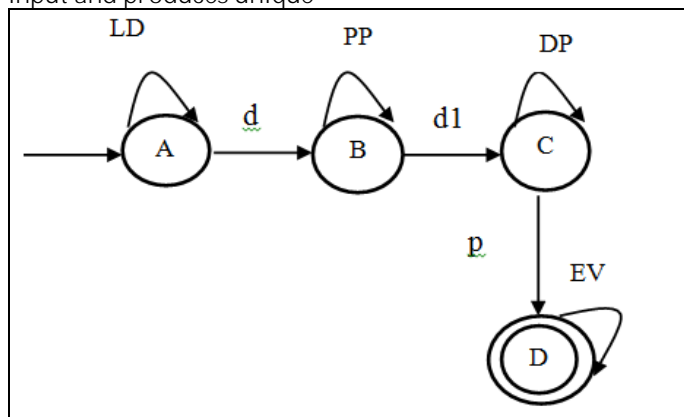


Fig -2: Deterministic Finite Automata
A deterministic finite automaton *M* is a 5-tuple, $(Q, \Sigma, \delta, q_0, F)$ consisting of-

- a finite set of states $(Q) = \{A, B, C, D\}$
- a finite set of input symbols called the alphabet $(\Sigma) = \{d, d1, p\}$

- a transition function $(\delta: Q \times \Sigma \rightarrow Q) = \{LD, PP, DL, EV\}$
- a start state $(q_0 \in Q) = \{A\}$
- a set of accept states $(F \subseteq Q) = \{D\}$

WHERE,

- d=document
- d1=document after removing stopwords and stemming
- p=patterns
- ep=Effective patterns
- LD=Loading Document
- PP=Preprocessing Loaded document
- DL=Deploying Patterns
- EV=Evolving Patterns

Derivation δ is defined in transition table 1.

Table -1: Derivation Table

States	d	d1	p
A	B	\emptyset	\emptyset
B	\emptyset	C	\emptyset
C	\emptyset	\emptyset	D
D	\emptyset	\emptyset	\emptyset

5.2 Set Theory

Let,

- D =Document set
- d =Single document
- PS(d) = Set of paragraphs in document d
- T =Set of terms
- D+ =Set of Positive document
- D- =Set of Negative Document
- P = Pattern set
- D = $\{d1, d2, d3, \dots, dm\}$
- PS(d) = $\{dp1, dp2, \dots, dpm\}$
- T = $\{t1, t2, \dots, tm\}$
- termset(pi) = $\{t1, t2, \dots, tm\}$
- $p1 \oplus p2 = \{(t, x1, x2) \mid (t, x1) \in p1, (t, x2) \in p2\} \cup \{(t, x) \mid (t, x) \in p1 \cup p2, \text{not}((t, _) \in p1 \cap p2)\}$
- P = $\{p1, p2, \dots, pm\}$

5.3 Multiplexer Logic

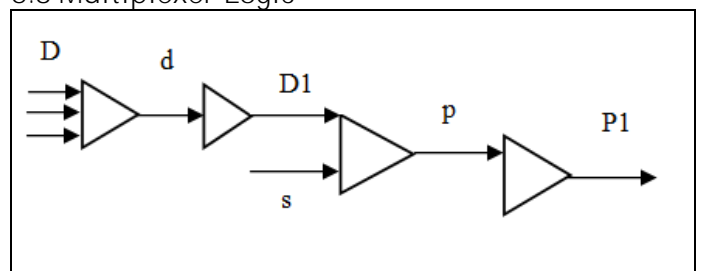


Fig -2: Multiplexer Logic

Where,

D=document set
 d= single document
 D1=document after preprocessing
 s=support
 p=extracted patterns
 P1=Final patters after removing noisy patterns

6 EXPERIMENTAL RESULTS

In this paper, we use text document collection as a input to evaluate the performance. We are applying several measures to compare the performance of system. Still system is not totally completed. Whatever outcomes of measures from this system are listed in table 2. Basically, in this results, we are used Breakeven point(b/p) measure, MAP,IAP measures.

6.1 Results of Existing System

Table -2: Existing System Results

Method	b/p	MAP	IAP
PTM	0.429	0.441	0.466
nGram	0.342	0.361	0.384
Rocchio	0.392	0.391	0.418
TFIDF	0.321	0.322	0.348
SVM	0.409	0.408	0.434

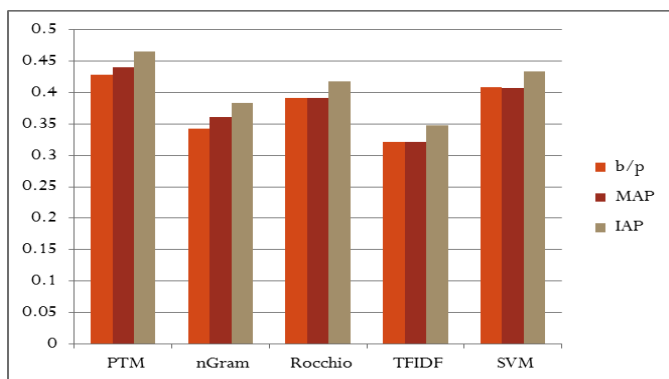


Chart -1: Existing System Results

6.2 Results of Proposed System

Table -3: Proposed System Results

Method	b/p	MAP	IAP
PTM(Proposed)	0.468	0.441	0.492
nGram	0.342	0.361	0.384
Rocchio	0.392	0.391	0.418
TFIDF	0.321	0.322	0.348
SVM	0.409	0.408	0.434

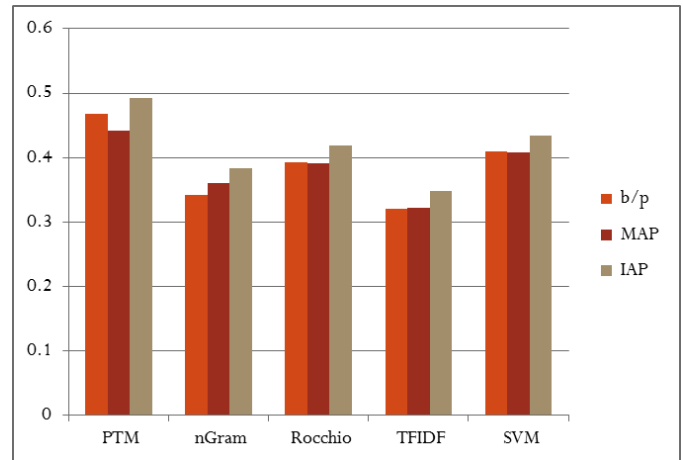


Chart -2: Proposed System Results

7. CONCLUSIONS

An innovative idea is presented for mining the textual data. Various methods have been developed to complete this task. Text mining approach provides discovering patterns. Main focus is on how to efficiently handle discovered pattern so that system get better performance by removing noisy patterns. Also for user point of view, this system provides the mechanism of clustering to make collection of related documents together.

REFERENCES

- [1] Ning Zhong, Yuefeng Li, Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining," , IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, January 2012.
- [2] Nitin Jindal and Bing Liu, "Identifying Comparative Sentences in Text Documents", University of Illinois at Chicago
- [3] Mrs.K. Mythili, and Mrs. K. Yasodha, "A Pattern Taxonomy Model with New Pattern Discovery Model for Text Mining", International Journal of Science and Applied Information Technology, Volume 1, No.3, July - August 2012
- [4] Deepshikha Patel, Monika Bhatnagar, "Mobile SMS Classification", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307 (Online), Volume-I, Issue-I, March 2011.
- [5] Ranveer Kaur, Shruti Aggarwal, "Techniques for Mining Text Documents", International Journal of Computer Applications (0975 - 8887),Volume 66- No.18, March 2013.

- [6] Atika Mustafa, Ali Akbar, and Ahmer Sultan, "Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization", International Journal of Multimedia and Ubiquitous Engineering Vol. 4, No. 2, April, 2009
- [7] Rashmi Agrawal, Mridula Batra, "A Detailed Study on Text Mining Techniques", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013.
- [8] Vishal gupta and Gurpreet S. Lehal , "A survey of text mining techniques and applications", *journal of emerging technologies in web intelligence*, 2009,pp.60-76.
- [9] Falguni N. Patel, Neha R. Soni," Text mining: A Brief survey", International Journal of Advanced Computer Research (ISSN (pri nt): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-4 Issue-6 December-2012.
- [10] N. Kanya and S. Geetha , "Information Extraction: A Text Mining Approach", IET-UK International Conference on Information and Comm. Technology in Electrical Sciences, IEEE(2007), Dr. M.G.R. University, Chennai, Tamil Nadu, India,1111- 1118.
- [11] <http://people.ischool.berkeley.edu/~hearst/text-mining.html>
- [12] <http://www.cis.upenn.edu/~ungar/KDD/text-mining.html>
- [13] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, (2005), "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg.
- [14] Dion H. Goh and Rebecca P. Ang," An introduction to association rule mining: An application in counselling and help seeking behaviour of adolescents", Journal of Behaviour Research Methods39 (2), Singapore, 259-266,2007.