Information Retrieval in Text Mining Using Pattern Based Approach

Minakshi R. Shinde[1], Prof. S. A. Kinariwala[2]

[1]Research Student, Marathwada Institute of Technology, Aurangabad, Maharashtra State, India

[2]Assistant Professor CSE Dept, Marathwada Institute of Technology, Aurangabad, Maharashtra state, India

Abstract: ***In text documents 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 an open research issue, especially in the text mining. So most existing text mining methods adopted term-based approaches but they all suffer from the problems of polysemy and synonymy. Polysemy is the word which giving the multiple meaning of word and synonymy is the word which giving the similar meaning of word. After some years, people have been adopted pattern based approaches should perform better than the term-based approaches. This paper with 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. In this paper user also getting the meaningful information without misinterpretation problem***

Keywords: Text mining, text classification, pattern deploying, pattern evolving, data mining

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. 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. 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 . 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. 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 . 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. 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.

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 pre-processing, pattern taxonomy model, pattern deploying process, evolving mechanisms. So that proposed system is dived into four modules that represent these processes.

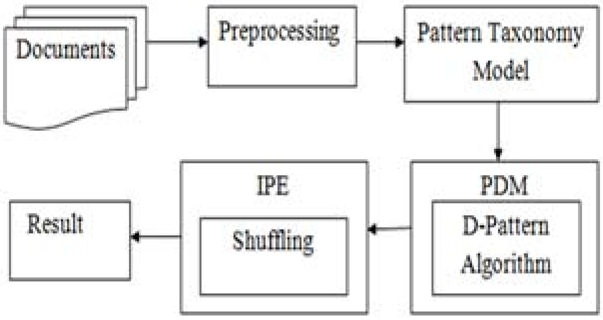

Figure 4.1: System Architecture

Proposed system divides the whole work into various stages v.i.z. pre-processing, 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.
Steps:
$DP= \emptyset$ ;
for each document $d$ € $D^+$ do
    let $PS(d)$ be the set of paragraphs in $d$;
    $SP$ =SPMining ($PS(d)$, $min\_sup$);
    d = ∅;
    for each pattern pi € $SP$ do
        $p=\{(t,1)|$ $t$ € $pi\}$;
        d = d ⊕ p ;
    end
    $DP=DP \cup \{d\}$;
end
$T=\{t|(t,f)$ € $p$ , $p$ € $DP\}$;
for each term $t$ € $T$ do
    $support\ (t)=0$;
end
for each d-pattern $p$ € $DP$ do
    for each $(t,w)$ € $\beta(p)$ do
        $support\ (t) = support\ (t) + w$;
    end
end
```

The pattern taxonomy model improves the semantic meaning of the discovered pattern by using the SPMining, which is 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 test process. After testing system's performance is evaluated using metrics such as precision, recall and f1-measures shows in equation. 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)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 $\mu$.
Output: a set of term-support pairs $np$.
Steps:
$np \leftarrow \emptyset$;
$threshold = Threshold\ (DP)$;
for each noise negative documents $nd$ € $D$ - do
    If $weight(nd) \geq threshold$ then $\Delta(nd) = \{p$ € $DP$ | $termset(p) \cap nd \neq \emptyset\}$;
    $NDP = \{\beta(p)$ $p$ € $DP\}$;
    Shuffling ($nd$,$\Delta(nd)$, $NDP$ , $\mu$ , $NDP$);
    for each $p$ € $NDP$ do
        $np \leftarrow np \oplus$ p;
    end
end
```

Algorithm 3: Shuffling

```
Input: a document d and a list of deployed patterns Δp.
Output: updated deployed patterns.
Steps:
For each deployed pattern $d$ in Δp do begin
    If termset($d$) ⊆d then // complete conflict offender
        Ω = Ω- {Ap}
    Else // partial conflict offender
        $offering = (1-\frac{1}{\mu}) \times \sum_{t \in termset(Ap)} \{t.weight \mid t \in d\}$
        $base = \sum_{t \in termset(Ap)} \{t.weight \mid t \in d\}$
        For each term $t$ in termset($Ap$) do begin
            If $t$ ϵ $d$ then // shrink offender weight
                $t.weight = \frac{1}{\mu} \times t.weight$
            else //shuffle weights
                t.weight = $t.weight \times \left(\frac{1+offering}{base}\right)$
            End if
        End for
    End if
End for
```

5. Mathematical Model

5.1 Deterministic Finite Automata (DFA)

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

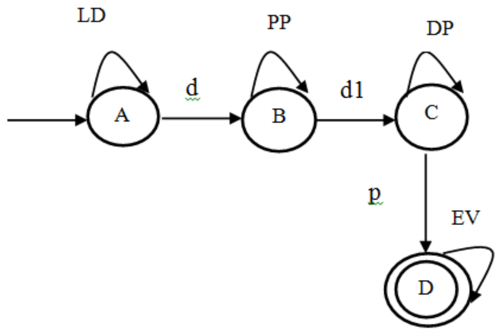

Figure 5.1: Deterministic Finite Automata

A deterministic finite automaton M is a 5-tuple, (Q,Σ, δ, q_0 ,F) consisting of:

- a finite set of states (Q)={A,B,C,D}
- a finite set of input symbols called the alphabet(Σ)={d,d1,p}
- a transition function (δ: Q x Σ->Q)={ LD,PP,DL,EV}
- a start state (q_0 ∈ Q)={A}
- a set of accept states (F ⊆ Q)={D}

Derivation δ is defined in transition Table 1.
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

Table 1: Derivation Table

States	d	d1	P
A	B	ϕ	ϕ
B	ϕ	C	ϕ
C	ϕ	ϕ	D
D	ϕ	ϕ	ϕ

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, ….., dm} PS(d)={dp1, dp2, ….., dpm}
T = {t1, t2, ….., tm},
termset (pi)= {t1, t2, ….., tm}
p1 ⊕ p2={ (t, x1, x2) | (t, x1) Є p1, (t, x2) Є p2} ∪{(t,x)|(t,x) Є p1∪p2, not((t, _) Є p1 ∩ p2) }
P = { p1, p2, ….., pm }

5.3 Multiplexer Logic

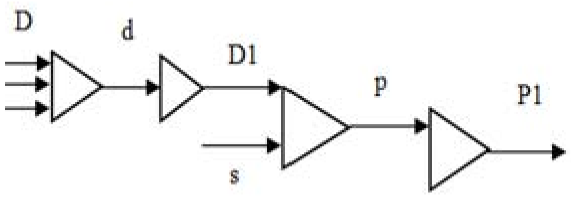

Figure 5.2: Multiplexer Logic

Where,
D=document set
d= single document
D1=document after pre-processing
s=support
P=extracted patterns
P1=Final patterns after removing noisy patterns.

6. Experimental Results

A popular text collection Reuters-21578 is used which has 21578 documents collected from the Reuters newswire. Among 90 categories, only the most popular 5 are used as shown in Table 2. Each category is employed as the positive examples class, and the rest as the negative examples class. This gives us 5 datasets.

Table 2: The most popular 5 categories on Reuters-21578 and their quantity

Categories	*Quantity*
Acq	537
Crude	571
Earn	2374
Money-fx	1208
Wheat	964

1) Evaluation Measure

In our experiments, we use the popular F1 score on the positive examples class as the evaluation measure. F1 score takes into account of both recall and precision. Precision, recall and F1 defined as:

$$Precision = \frac{\#\, of \;\; relevant\;\, terms}{\#\, of \;\;\; retrieved\;\; terms}$$

$$Recall = \frac{\#\, of \;\; relevant\;\, terms}{\#\, of \;\;\; terms\;\; computed}$$

$$F_1 = \frac{2 * precision * recall}{precision + recall}$$

For evaluating performance average across categories, macro-average is used. Macro-averaged performance scores are determined by first computing the performance measures per category and then averaging those to compute the global means. We use macro-averaging.

2) Experimental Results:

We have implemented this information retrieval technique using reuters-21578 collection taking five popular categories. Table 3 shows the results of PDM method the relevant terms are found out from positive document set. Table 4 shows the results of IPE method the relevant terms are found out from positive document set.

Table 3: Performance evaluation for PDM method

Topic	*Precision*	*Recall*	*F-measure*
Acq	0.85	0.43	0.57
Crude	0.69	0.37	0.48
Earn	0.45	0.21	0.29
Money-fx	0.7	0.4	0.50
Wheat	0.69	0.32	0.44

Table 4: Performance evaluation for IPE method

Topic	*Precision*	*Recall*	*F-measure*
Acq	0.88	0.76	0.81
Crude	0.54	0.4	0.46
Earn	0.98	0.45	0.61
Money-fx	0.5	0.4	0.45
Wheat	0.71	0.5	0.58

In PDM macro average precision score is 67% , recall score is 34% and f-measure score is 45% terms retrieved to know the system's overall performance across sets of data. In IPE macro average precision score is 72%, recall score is 50% and f-measure score is 58% terms retrieved to know the system's overall performance across sets of data. Performance evaluation show that the pattern Evolving is superior to pattern deployment data mining-based method.

7. Conclusion

Data mining techniques provides pattern mining methods but to use these patterns and update to solve misinterpretation and low frequency problem is achieved in this approach. Knowledge discovery with PDM and IPEvolving have been proposed to overcome the misinterpretation & low frequency problem. An effective knowledge discovery system is implemented using three main steps: (1) discovering useful patterns by sequential pattern mining algorithm (2) Using D-pattern discovery, term support evolution is done. (3) IPEvloving is used to reduce the influence of noisy terms. The experimental results show that the proposed model improves the performance of finding the accurate knowledge from the text data.

8. Acknowledgements

The authors express gratitude to Principal, Head of Department (CSE) Marathwada Institute of Technology College of Engineering, Aurangabad, and Maharashtra India. They also express their sincere thanks all the faculty members of CSE Department MIT College of Engineering, Aurangabad, and Maharashtra, India for their constant support and enthusiasm.

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.

Author Profile

Minakshi R.Shinde received the B.E degree in Information Technology from HI-TECH Institute of Technology from Aurangabad in 2011 At present appearing the M.E degree in Computer Science and Engineering department at Marathwada Institute of Technology, Aurangabad. Her research interest in Information Retrieval in Text Mining Using Pattern Based Approach