A NOVEL APPROACH FOR TEXT EXTRACTION USING EFFECTIVE PATTERN MATCHING TECHNIQUE

Bhaskaran Shankaran[1], Milindkumar Patil[2], Shankar Suryawanshi[3], Sandesh Mandhane[4], S.S.Raskar[5]

[1]*Student [BE], Computer, MES College Of Engineering, Pune, Maharashtra, India*
[2]*Student [BE], Computer, MES College Of Engineering, Pune, Maharashtra, India*
[3]*Student [BE], Computer, MES College Of Engineering, Pune, Maharashtra, India*
[4]*Student [BE], Computer, MES College Of Engineering, Pune, Maharashtra, India*
[5]*Assistant Professor, Computer, MES College Of Engineering, Pune, Maharashtra, India*

Abstract

There are many data mining techniques have been proposed for mining useful patterns from documents. Still, how to effectively use and update discovered patterns is open for future research , especially in the field of text mining. As most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy(words have multiple meanings) and synonymy(multiple words have same meaning). People have held hypothesis that pattern-based approaches should perform better than the term-based, but many experiments does not support this hypothesis. This paper presents an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern matching, to improve the effective use of discovered patterns.

Keywords: *Pattern Mining, Pattern Taxonomy Model, Inner Pattern Evolving, TF-IDF, NLP etc.*

----------***----------

1. INTRODUCTION

We are deluged by data—scientific data, medical data, financial data, marketing data and many more resources. People have no time to look at all these data and are interested only in those which are useful for them in their work or business. Human attention has become a precious resource. So, we must find ways to automatically analyze classify and summarize the data for easier access of the required data. This is one of the most active and exciting areas of the database research community.

In this paper we have discussed the design and implementation of effective pattern matching technique for text mining. Text mining is the discovery of interesting knowledge in text documents. It is a challenging task to find accurate requirement (or features) in text documents to help users to find what they want. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. The large number of patterns generated can be effectively used and can update these patterns using various data mining approaches. The purpose of this project is to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Further sections are described as follows.

In section 2 the Existing System is briefed, in section 3 the Proposed System is seen. The other coming sections describe the Objectives and Scope, other related work, conclusion and acknowledgement respectively.

1.1 Text Mining

Text mining, also referred as *text data mining*, refers to the process of obtaining high-quality information from processed text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text, deriving patterns and finally evaluating and interpreting the output. Typical text mining tasks include text categorization, text clustering, concept, document summarization. Text analysis consists of information retrieval, lexical analysis to study word frequency distributions, pattern recognition, and information extraction. The overarching goal is to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods by using the proposed algorithm TF-IDF [Term Frequency-Inverse Document Frequency].

1.2 Pattern Mining

"Pattern mining" is a data mining method that involves finding existing patterns in the database. In this context *patterns* are referred to association rules. The original motivation for searching frequent patterns came from the desire to analyze transaction data of large business organizations; educational institutes etc. in order to examine the statistics in terms of transactions done. For example, an association rule admission⇒ books (80%)" states that four out of five students that took admission to a particular school also bought books from the same institute.

Some types of Pattern Mining are:
Association Rule Mining: It is a method for discovering interesting relations between variables in large databases. In contrast with sequence mining, association rule mining does not consider the order of items either within a transaction or across transactions.

Sequential Pattern mining: It is a topic of data mining which is concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. It is usually presumed that the values are discrete in this case. Thus Sequential pattern mining is a special case of structured data mining.

1.3 TF-IDF Algorithm

The algorithm used for obtaining and discovering unique patterns and eliminating noisy patterns is TF-IDF algorithm.TF means Term Frequency and IDF means Inverse Document Frequency.

This is often used as a weighting factor in information retrieval. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word, which signifies that some words appear more frequently in general. Variations of the TF-IDF weighting scheme are often used as a central tool in ranking a document's relevance. TF-IDF can be successfully used for stop-words filtering in various subject fields including text summarization and classification. In the case of the **term frequency** tf(*t*,*d*), the simplest choice is to use the *raw frequency* of a term in a document, i.e. the number of times that term *t* occurs in document *d*. If we denote the raw frequency of *t* by f(*t*,*d*), then the simple tf scheme is tf(*t*,*d*) = f(*t*,*d*).

$$\mathrm{tf}(t,d) = 0.5 + \frac{0.5 \times \mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$

The **inverse document frequency** is a measure of how much information the word provides, that is, whether the term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term, and then taking the logarithm of that quotient.

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

with N: total number of documents in the corpus

$|\{d \in D : t \in d\}|$: number of documents where the term $t$appears.If the term is not in the corpus, this will lead to a division-by-zero. It is therefore common to adjust the denominator to $1 + |\{d \in D : t \in d\}|$.

2. EXISTING SYSTEM

Existing is used to term-based approach to extracting the text. Term-based ontology methods are providing some text representations. Pattern evolution technique is used to improve the performance of term-based approach.

E.g.: Hierarchical is used to determine synonymy and hyponymy relations between keywords.

Demerits of Existing System:

•The term-based approach is suffered from the problems of polysemy and synonymy.
•A term with higher (tf*idf) value could be meaningless in some d-patterns (some important parts in documents).

3. PROPOSED SYSTEM

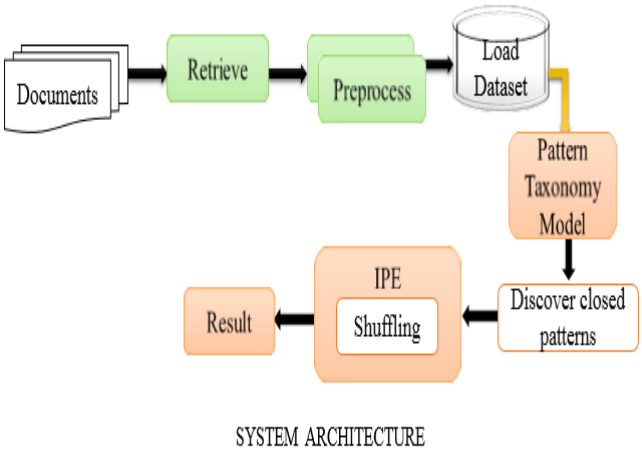
In the proposed system an effective pattern discovery technique, is discovered which specifies the patterns and then evaluates term weights according to the distribution of terms in the discovered patterns. It also solves misinterpretation problem, considers the influence of patterns from the negative training examples to find ambiguous (noisy) patterns and tries to reduce their influence for the low-frequency problem. The process of updating ambiguous patterns can be referred as pattern evolution.

3.1 System Architecture

The System Architecture is shown. It consist of the following modules:

A. **Document:** This is the input which is given by the end user. It could either be a text or a word or even a PDF document.

B. **Preprocess:** After the document is retrived into the preprocessor it performs two operations i.e to identify stop words like is, for, the ,a etc. and to perform stemming process i.e if we search for a word 'fill' then it produces related outputs like filling , filled etc. Thus it also generates 'ing' and 'ed' words.



SYSTEM ARCHITECTURE

Fig -1: System Architecture

After the above two process the dataset is loaded into the database.

C. **Pattern Taxonomy Model**: In this model the text paragraphs are considered as input and it evaluates line by line identifying frequent and closed patterns.

D. IPE Shuffling: IPE stands for Inner Pattern Evolving. It is used for checking if any ambiguous patterns are present and if found they are removed.

Thus the main purpose of this project is to first find frequent sequential pattern, compute closed sequential patterns, reshuffle the patterns to eliminate noisy (ambiguous) patterns using Inner Pattern Evolution and produce the necessary result.

3.2 Algorithmic Strategy

Divide and conquer: A divide and conquer strategy works by recursively breaking down a problem into two or more sub-problems of the same or related type, until these become simple enough to be solved directly. The solutions to the sub-problems are then combined to give a solution to the original problem. In the proposed system we are splitting multiple documents into paragraphs and using the paragraph fordoing further calculations. Along with divide and conquer strategy dynamic programming strategy is to mine the text from documents after removal of stop words.

Pseudo Code
Input - Text, Word, Pdf Documents
Ouput - Offsets of each term
actualOffset = 0s
EndOfLineOffset = 0
termArray[100]
for each d in D+ do
//Line ends with \n
for each Line in d do
Line = Remove Stop Words (Line)
Line = Streaming (Line)
termArray = Line.split (" ")
for each term in term Array do
UpdateTermInDB(term)
end
end
end
foreach d in D+ do
foreach Line in d do
if Line contains "$$" Then
//This means more than one data set(files) selected for text mining need to set EndOfLineOffset to 0
EndOfLineOffset = 0
end if
termArray = GetAllTermsFromDB(Line)
foreach term in termArray do
actualOffset = GetTheOffsetFromREGX(term,Line) + EndOfLineOffsetsUpdateTheOffsetInDB(actualOffset,term, Line)
end
EndOfLineOffset = EndOfLineOffset + Line.length() + 2 // +2 is for \r\n characters
End

4. OBJECTIVE AND SCOPE

The Objective and Scope of the proposed system are as follows.

4.1 Objective

To design and implement effective pattern search for text mining.Text mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining, and closed pattern mining. With a large number of patterns generated by using data mining approaches, how to effectively use and update these patterns.

4.2 Scope

The scope of our project is presently specific. There will be a single user at any given time using the system. He/she may use the system for searching particular book from database by using keywords like cloud, C, C++, etc. as input. This system will speed up the searching mechanism and also provide a different angle to search engines. Thus the main aim of our project is to satisfy the following features

- The purpose of this project is to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information.
- Able to solve the problem of the text mining on polysemy and synonymy effectively.

5. RELATED WORK

5.1 Pattern Discovery for Text Mining Using Pattern Taxonomy

In this paper the advantage is that it is useful for developing efficient mining algorithm for discovering pattern from large data collection.But the major disadvantage is that it generated large number of pattern not knowing to effectively exploit these pattern which is still an open research issue.

5.2 A Text Mining Approach towards Knowledge Management Applications

Although it was an effective pattern discovery technique it kept losing all the word order information and only retaining the frequency of the terms in the document.

6. PLATFORM

6.1 User Interface

The user interface includes Graphic User Interface (GUI) where the user enters the keyword to be searched for as an input. The output includes all the frequent patterns matching the entered keyword result & the device will generate the list of

related documents containing that particular keyword eliminating all the stop words like is, the, a etc.

6.2 Hardware Requirements

The minimum Hardware requirements for a PC are:

1. Operating System as Windows-7 or 8 having 2 or 4 GB RAM.
2. Hard Disk capacity of 40 G.B or higher.

6.3 Software Requirements

1. The language used to code the system is Java JDK 1.7.0 & JRE 6 or 7.
2. Eclipse or Net Beans 8.0 Version as programming is totally Java based.
3. For system designing the Software's required would be Star UML 5.0.

7. CONCLUSION

There are many data mining techniques that have been proposed for mining useful patterns from documents previously. These techniques include association rule of mining, sequential pattern mining, maximum, frequent item set mining, pattern mining, and closed pattern mining. It seems that the discovered knowledge (or patterns) in the field of text mining is difficult and ineffective. The reason is there is a very long pattern with high specificity lack in support (i.e., the low-frequency problem). And we argue that not all frequent short patterns are useful. Thus, misinterpretations of patterns derived from data mining techniques organize the weak performance of the system. The proposed technique uses four processes, pattern deploying and pattern evolving, shuffling, offset to refine the discovered patterns in text documents. These experimental results show that the proposed model outperforms pure data mining-based methods and the concept based model and the high-level performance of the system can be achieved. Security is enabled as certain features and functionalities are restricted to the registered users only. The system aims to respond to the user as fast as possible. The system is adaptable to the future changes if any are deployed in order to meet the technological advancements.

REFERENCES

[1]. Ning Zhong, Yuefeng Li, and Sheng-Tang Wu,"Effective Pattern Discovery for Text Mining", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL.24,NO. 1, JANUARY 2012.

[2]. Miss Dipti S.Charjan, Prof. Mukesh A.Pund, "Pattern Discovery For Text Mining Using Pattern Taxonomy", International Journal of Engineering Trends and Technology (IJETT) – Volume 4 Issue 10- October 2013.

[3]. Kavitha Murugeshan, Neeraj RK "Discovering Patterns to Produce Effective Output through Text Mining Using Naïve Bayesian Algorithm" IJITEE ISSN: 2278-3075, Volume-2, Issue-6, May 2013

[4]. S.Shehata, F.Karray and M.Kamel,"A Concept-Based Model for Enhancing Text Categorization," Proc. 13th Int'l Conf. Knowledge Discovery and Data Mining (KDD '07), pp. 629-637, 2007.

ACKNOWLEDGEMENTS

We would like to express a deep sense of gratitude and appreciation to all those who helped us in the completion of this paper. A special thanks to our project guide, **Prof. S.S Raskar**.Last but not the least, we would like to thank our friends and family for their valuable support.

BIOGRAPHIES

Bhaskaran Shankaran is with Modern Education Society's College of Engineering, Pune-01. He is currently pursuing BE in Computer, Savitribai Phule University of Pune.

Milindkumar Patil is with Modern Education Society's College of Engineering, Pune-01. He is currently pursuing BE in Computer, Savitribai Phule University of Pune.

Shankar Suryawanshi is with Modern Education Society's College of Engineering, Pune-01. He is currently pursuing BE in Computer, Savitribai Phule University of Pune.

Sandesh Mandhane is with Modern Education Society's College of Engineering, Pune-01. He is currently pursuing BE in Computer, Savitribai Phule University of Pune.