D-Pattern Algorithm for Text Mining

A. D. Khade[1] A. B. Karche[2] D. S. Jadhav[3] M. V. Vaghpatil[4] A. S. Zore[5]

Abstract---We know that multiple data mining methods have been developed for finding useful patterns in contents like PDF files, text files. Current paper addresses the problem of making text mining results more effective to humanities scholars, journalists, intelligence analysts, and other researchers. To use effective and bring to up to date discovered patterns is still an open research task, especially in the domain of text mining. Text mining is the finding of very interesting knowledge (or features) in the text documents. It is a very difficult to find exact knowledge (or features) in text documents to help users what they actually want. This paper represent efficient mining algorithm to find particular patterns.

Key words: d-pattern algorithm; pattern mining; pattern taxonomy; semantic algorithm

I. INTRODUCTION

Very fast growth of digital data in a recent years, knowledge discovery as well as data mining has robust importance. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. So a many patterns generated by using data mining techniques, how to perfectly use and update these patterns is a research task. In this paper, we focus on the development of a knowledge discovery model to effectively use and update the discovered patterns and apply it to the field of text mining.

Our project involves text mining technique. So it is need to understand text mining concept.

Text Mining: Text mining is nothing but finding the useful information from the text documents. Text mining techniques used to solve business problems is called text analytics.

II. PATTERN CONCEPT

Text mining is the discovery of useful and interesting knowledge in text documents. It is a very hard job to find out exact information in text documents which perhaps helps to user's requirement. In the starting phase, Information Retrieval (IR) uses 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. However, term based methods have a problems of polysemy and synonymy, where polysemy means a word has multiple meanings, and synonymy is multiple words has same meaning . Although phrases are less ambiguous and finer difference than individual terms, which leads to discouraging performance contain:

1) Phrases have low frequency of occurrence,
2) They have inferior statistical properties to terms, and
3) There are large numbers of occur of phrase and noisy phrases among them

III. WORKING

A d-pattern mining technique is discovered. It evaluates specificities of patterns and then evaluates term-weights according to the distribution of terms in the discovered patterns. It solves Misinterpretation Problem. For example, term "LIB" may have more weight than "JDK" in a certain text file; but we all know that term "JDK" is more specific

than term "LIB" for describing "Java Programming Language"; and term "LIB" is more general than term "JDK" because term "LIB" is also frequently used in C and C++.Therefore, it is not simple for evaluating the weights of the terms depending upon their distributions in documents. In order to solve the above problem, current paper uses a d-pattern mining technique, which calculates specificities of patterns first 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 impacts of patterns from the negative training examples to find ambiguous patterns and try to reduce their impact for the low-frequency problem. The process of improving 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.

In General there are two phases:

Training and Testing:-

Training: In training phase the d-patterns in positive documents (D) based on a min sup are found, and evaluates term supports by deploying patterns to terms.

Testing: In Testing Phase to revise term supports using noise negative documents in D based on an experimental coefficient. The incoming documents then can be sorted based on these weights.

1) Datasets:-Dataset is collection of the data which present in tabular form i.e. we can represent the data in row & column wise format.

IV. LITRATURE SURVEY

Existing System	Disadvanteges
1.Term Based	Polysemy, Synonymy
2.Phrase Based	Low Frequancy

V. SYSTEM ARCHITECTURE

The proposed n architecture is shown in Figure 1.

This architecture shows that step by step execution of our project. The first step is to load documents in our dataset. The next step is to remove stop word and perform

text steaming. We removed this stop word and text steaming with the help of NLP (natural language process).

There are 5 sub modules of System Architecture.

1) Load the documents
2) Preprocessing of text
3) Splitting of Paragraph
4) Deploying the pattern
5) Testing the pattern

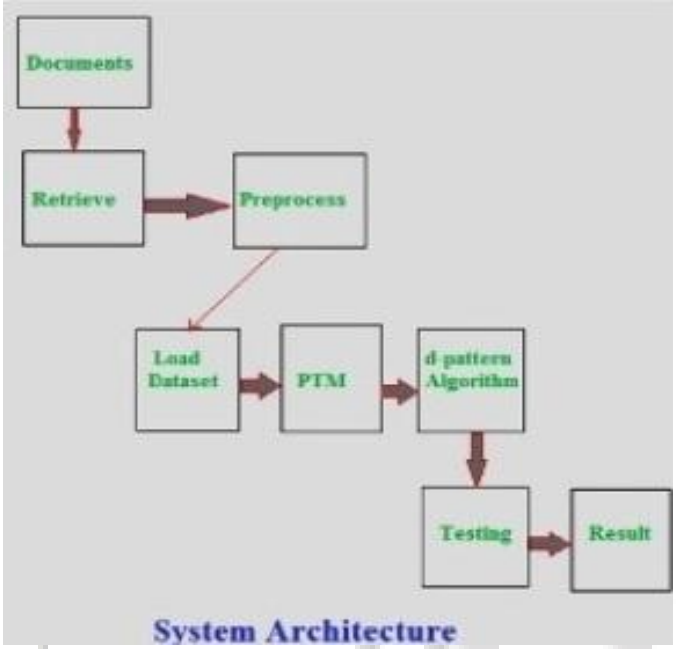
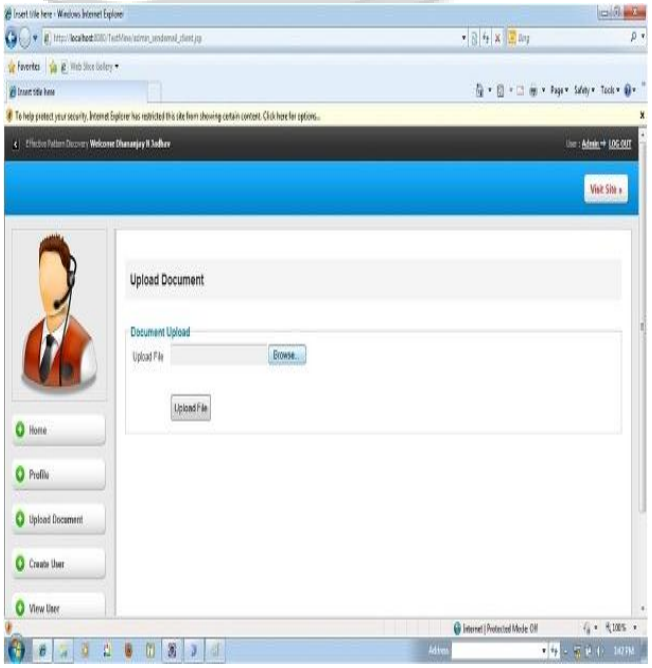


Fig. 1: System Architecure

1) Loading documents

In this module, we load the documents as per our need. Then user can retrieve any one of them documents. This document is given to next process which is nothing but preprocessing



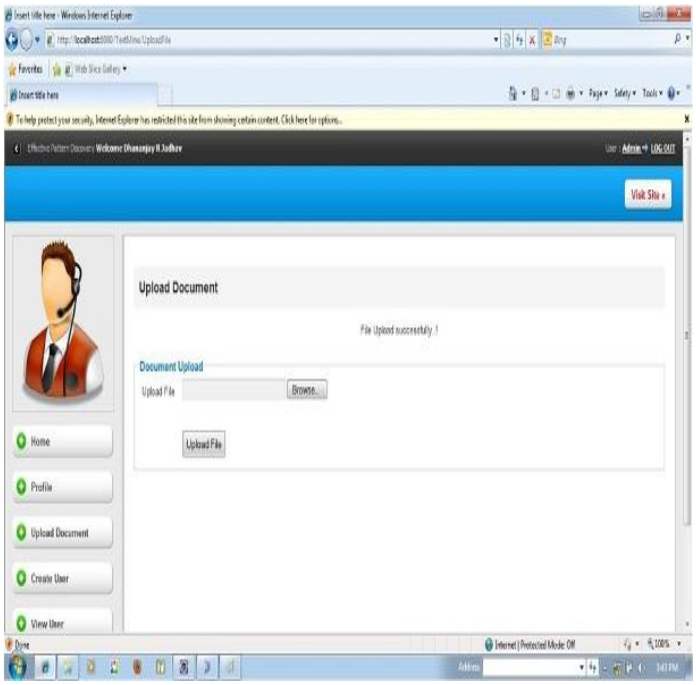


Fig. 2: Successfully Loaded Document

2) Text Preprocessing

The retrieved document preprocessing is done in module.

With help of two sub processes such as,

a) Stop words removal
b) Text stemming

Stop words are words which are filtered out prior to, or after, processing of natural language data. Stemming is the process for reducing inflected (or sometimes derived) words to their stem base or root form. It generally a written word forms.

3) Pattern taxonomy process

In this module, the documents are split into paragraphs and each paragraph is considered to beach document, which leads to the terms which can be extracted from set of positive documents.

Where d=document;

m= set of paragraph; s=keyword;

Paragraph	Terms
dm1	s1, s2
dm2	s3, s4 , s6
dm3	s3 , s4, s5, s6
dm4	s3, s4, s5, s6
dm5	s1 , s2, s6, s7
dm6	s1 , s2, s6 , s7

Table. 1: A Set of Paragraph

Frequent Pattern	Covering sets
{ s3, s4, s6}	{dm2,dm3,dm4}
{ s3 , s4 }	{dm2,dm3,dm4}
{ s3 , s6}	{dm2,dm3,dm4}
{ s4 , s6}	{dm2,dm3,dm4}
{ s3 }	{dm2,dm3,dm4}
{ s4 }	{dm2,dm3,dm4}
{ s1 , s2 }	{dm1,dm5,dm6}
{ s1 }	{dm1,dm5,dm6}
{ s2 }	{dm1,dm5,dm6}

{ s6}	{dm2,dm3, dm4,dm5,dm6}

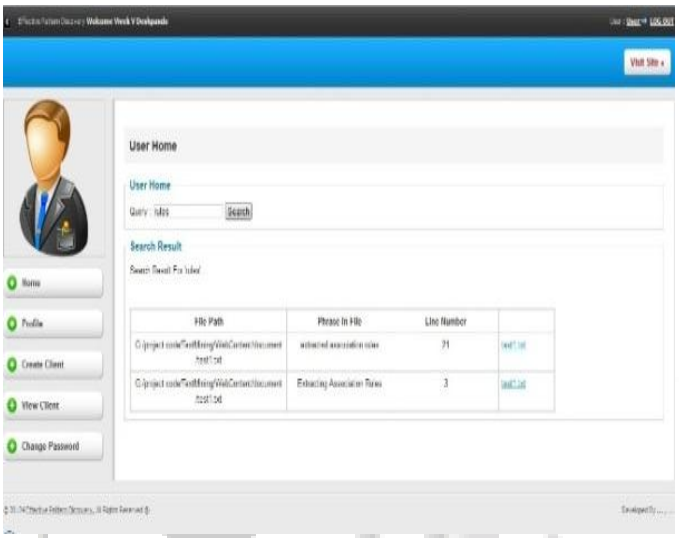
Table 2: Frequent Pattern and Covering set

4) Pattern deploying

The discovered patterns are summarized. The d-pattern algorithm is used to discover all patterns in positive documents are composed. The term supports are calculated by all terms in d-pattern. Term support means weight of the term is evaluated.

5) Pattern Testing

In this module used to identify the noisy patterns in documents. Sometimes, system falsely identified negative document as a positive. So, noise is occurred in positive document. The noised pattern named as offender. In positive documents, the reshuffle process is done in case of partial conflict offender.



VI. Expected Result

To focus on the development of knowledge discovery model to effectively use and update the discovered patterns and apply it to the field of text mining. Technology will help in fast finding of text. There is efficient use of text mining. Finding searching pattern with their location in effectively.Processing and Multilingual Aspects are present in system.

VII. Technical Specification

Software**:**
1. Dataset used-RCV1
2. Coding Language JAVA
3. Tools and Databases used-My-SQL
4. Operating System Windows XP /7

Hardware:
1. Ram: 512 MB. (Min)
2. Hard Disk: 40 GB. (Min)
3. System: Pentium IV 2.4 GHz.

Acknowledgment

We would like to sincerely thank Mr. A.S. Zore, our mentor (Lecturer, MMIT, Lohgaon), for his valuable support and Encouragement.

References

[1] J. Han and K.C.-C. Chang, "Data Mining for Web Intelligence,"
[2] A. Maedche, Ontology Learning for the Semantic Web. Kluwer Academic, 2003
[3] Y. Yang, "rsAn Evaluation of Statistical Approaches to Text Categorization," lnformation Retrieval.
[4] "Interpretations of Association Rules by Granular Computing," By Y. Li and N. Zhong.
[5] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrival, Addison Wesley, 1999
[6] Data set RCV1. http://www.RCV1dataset.com/home/
[7] Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections by H. Ahonen, O Heinonen, M. Klemettinen, and A.I. Verkamo.
[8] K. Aas and L. Eikvil, "Text Categorization: A Survey ," Technical Report Raport NR 941, Norwegian Computing Center, 1999

.

