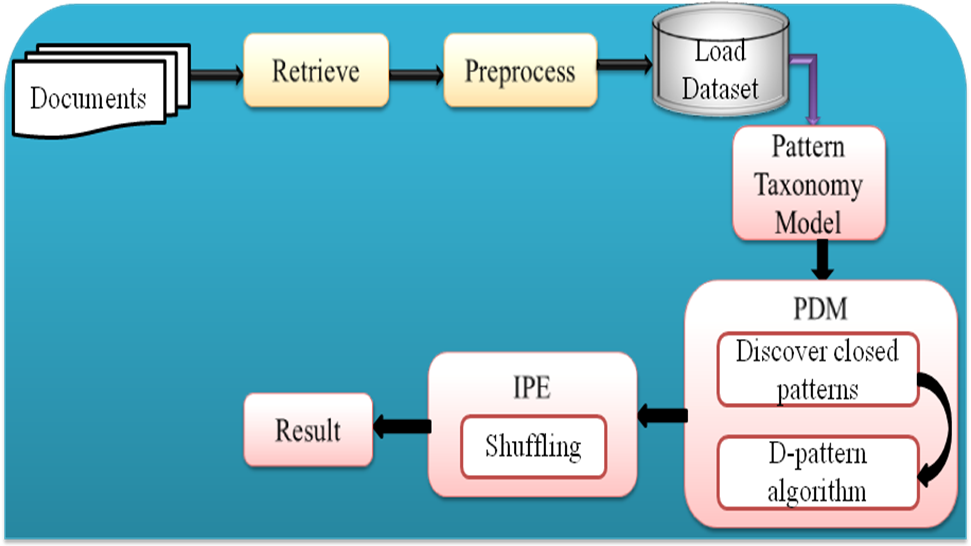
**Effective Pattern Discovery for Text Mining**

**ABSTRACT:**

Many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns is still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support this hypothesis. This paper presents an 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. Substantial

**SYSTEM ARCHITECTURE:**



**EXISTING SYSTEM:**

* Existing is used to term-based approach to extracting the text.
* Term-based ontology methods are providing some text representations.
* E.g.: Hierarchical is used to determine synonymy and hyponymy relations between keywords.
* Pattern evolution technique is used to improve the performance of term-based approach.

**DISADVANTAGES 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).

**PROPOSED SYSTEM:**

* An effective pattern discovery technique, is discovered
* Evaluates specificities of patterns and then evaluates term weights according to the distribution of terms in the discovered patterns
* 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.
* 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
* In training phase the d-patterns in positive documents (Dþ) based on a min sup are found, and evaluates term supports by deploying dpatterns to terms
* 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.

**ADVANTAGES OF PROPOSED SYSTEM:**

* The proposed approach is used to improve the accuracy of evaluating term weights.
* Because, the discovered patterns are more specific than whole documents.
* To avoiding the issues of phrase-based approach to using the pattern-based approach.
* Pattern mining techniques can be used to find various text patterns.

**LIST OF MODULES:**

1. Loading document
2. Text Preprocessing
3. Pattern taxonomy process
4. Pattern deploying
5. Pattern evolving

**MODULES DESCRIPTION:**

1. **Loading document**

* In this module, to load the list of all documents.
* The user to retrieve one of the documents.
* This document is given to next process.
* That process is preprocessing.

1. **Text Preprocessing**

* The retrieved document preprocessing is done in module.
* There are two types of process is done.
* 1) stop words removal 2)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.

1. **Pattern taxonomy process**

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

1. **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.

1. **Pattern evolving**

* 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.
* If partial conflict offender contains in positive documents, the reshuffle process is applied.

# SYSTEM CONFIGURATION:-

# HARDWARE REQUIREMENTS:-

# Processor -Pentium –III

* Speed - 1.1 Ghz
* RAM - 256 MB(min)
* Hard Disk - 20 GB
* Floppy Drive - 1.44 MB
* Key Board - Standard Windows Keyboard
* Mouse - Two or Three Button Mouse
* Monitor - SVGA

# SOFTWARE REQUIREMENTS:-

* Operating System : Windows95/98/2000/XP
* Front End : Java
* TOOL : Netbeans IDE

**REFERENCE:**

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.**