**ABSTRACT**

Data mining is an essential step in the process of knowledge discovery in databases. With a large number of patterns generated by using data mining approaches, how to effectively use and update these patterns is still an open research issue. 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. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. More text sequences are being generated in various forms. To discover valuable knowledge from a text sequence, the first step is usually to extract topics from the sequence with both semantic and temporal information, which are described by two distributions, respectively: a word distribution describing the semantics of the topic and a time distribution describing the topic’s intensity over time. Plagiarism with Text mining attempts to discover new, previously unknown information by applying techniques from natural language processing and data mining. Our system find exact plagiarism count for different documents.

1. **INTRODUCTION**
   1. **Motivation**

Text mining is a method that discovers interesting information in text documents. Text mining is gaining attention because of its automatically discovering technique of knowledge assets buried in unstructured text. It is a challenge to find accurate feature in text documents to help users to find what they want. Developing efficient feature extraction algorithms is highly needed to deal with high-dimensional data sets. In text mining techniques, the frequency of a word or phrase is observed to gain the importance of the term in the document. And text clustering helps in selecting important feature, which has critical effect on output of clustering algorithm.

This work performs multiple levels of automatic text analysis to identify the specific phase from the given documents. It also improves the performance of the pattern mining on text documents feature extraction method is used for information extraction from given text document. Here a work is proposed that will generate summary of the text documents that is research papers given by user on the phase given by the user.

* 1. **Aim**

Aim of this work is to study some methods along with their respective work contribution and to propose an improved pattern mining technique for information retrieval of text documents using feature extraction and text mining and find out plagiarism count.

* 1. **Objective**

The objectives are as follows:

1. Using old synopsis only, generate plagiarism count of different type of paper which may or may not dependent on particular domain.

2. The aim is to insert any no. of paper or synopsis to summarize data.

3. Plagiarism count can be done with sentence based,.

* 1. **Scope**

The scope of improved pattern mining technique for information retrieval of text documents using feature extraction and text mining will be the research papers that provided by the user for plagiarism analysis.

Here a work is proposed that will generate plagiarism count of the text documents that is research papers given by user on the phase given by the user.

1. **LITERATURE SURVEY**
   1. **Background History**

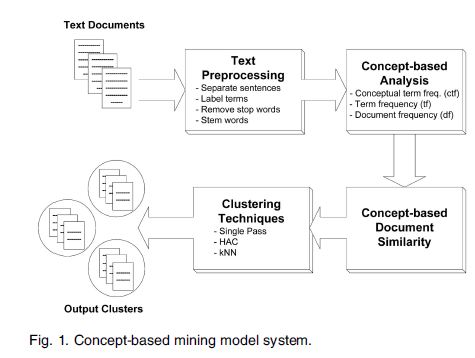
Concept-based mining model proposed by S.Shehata et. al., consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure. The objective behind the concept-based analysis task is to achieve an accurate analysis of concepts on the sentence, document, and corpus levels rather than a single-term analysis on the document only. A new concept based mining model composed of four components, is proposed to improve the text clustering quality. This method has drawback that it cannot work on web documents [1]. Ontology-based text mining method (OTMM) to cluster research proposals was a hybrid method for grouping Chinese research proposals for project selection was proposed by Jian Ma et al. It uses text-mining, multilingual ontology, optimization, and statistical analysis techniques to cluster research proposals based on their similarities. The proposed OTMM was used together with statistical method and optimization models. The proposed method promotes the efficiency in the proposal grouping process [2]. Ontology-based text mining methodology constructs the D-matrices by automatically mining the unstructured repair verbatim data collected during fault diagnosis was given by D.Rajpathak et al. The method is implemented as a prototype tool and validated by using real-life data collected from the automobile domain. The text-driven D-matrix of huge size helped the engineering users to perform the root-cause analysis accurately

* 1. **Existing System**

**VARIOUS EXISTING METHODS**

Many text mining methods has been implemented over the last decades. There are different methodologies that are implemented for mining text document. They are as follows:

**Concept-based mining model:** It consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure, as depicted in Fig. 1. Figure shows the methodology of this model[1].



**Feature clustering algorithm:** This deal with following issues.

*1 Self-Constructing Clustering:* This clustering algorithm is an incremental, self-constructing learning approach. Word patterns are considered one by one.

*2 Feature Extractions:* By applying this algorithm, word patterns have been grouped into clusters, and words in the feature vector W are also clustered accordingly. For one cluster have one extracted feature.

*3 Text Classifications:* The similarity threshold is applied to clustering algorithm. Assuming k clusters are obtained for the words in the feature vector W. Then finds the weighting matrix T and convert D to D0. Using D0 as training data, a classifier based on support vector machines (SVM) is built[2].

**Effective pattern discovery:** It is a technique that first calculates discovered specificities 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 patterns and try to reduce their influence for the low-frequency problem. This technique can improve accuracy of evaluating term weights because discovered patterns are more specific than whole documents[4].

**Topic mining from multiple asynchronous text sequences:** The objective function is given which is to maximize the likelihood estimation subject to certain constraints. The key idea of this approach is to utilize the semantic and temporal correlation among sequences and to build up a mutual reinforcement process. The outline of the algorithm is:

Step 1: extract common topics from them.

Step 2: Synchronize the time stamps of all documents by matching them to most related topics.

*Topic Extraction:* The current time stamps of all sequences are already synchronous and extract common topics from them. *Time Synchronization:* Once the common topics are extracted, adjust their time stamps to synchronize the sequences[5].

**Ontology-based Text Mining Method:** OTMM uses statistical and optimization models and have four phases. First, a research ontology containing the projects of latest five years is organized according to keywords. Then, new research proposal is classified to its area using a sorting algorithm. Next, the new proposals in each discipline are clustered using a self-organized mapping algorithm. Finally, if the number of proposals in each cluster is still very large, they will be further decomposed into subgroups[6].

**Learning Discriminative phrase pattern mining**: It is an unsupervised method for feature learning, so it is difficult to optimize for discrimination. Mutual information is the reduction of uncertainty in a random variable after observing another random variable. This uses: i) the mutual information of the phrase pattern to determine if a phrase pattern is discriminative, and ii) the upper-bound for mutual information against the threshold to determine if any extensions of the phrase pattern may be discriminative[8].

**Low Rank Shared Concept:** LRSC domain methodology is as follows:

1. This discovers low-rank shared concept space where the empirical loss on the labeled data and the distribution gap between source and target domain are jointly minimized.

2. It can kernelize method to generalize the model by making use of the powerful kernel functions.

3. Analyzes the expected error evaluated by common loss functions in the target domain under the empirical risk minimization framework, the error bound can be controlled by the expected loss in the source domain and the embedded distribution gap.

4. Domain adaptation method is capable of considering multiple classes and their interactions simultaneously[9].

**Ontology-based text mining method for Text-driven D-matrix:** It automatically constructs and updates mining thousand of repair verbatim collected during the diagnosis episodes. In this, it first constructs the fault diagnosis ontology consisting of concepts and relationships commonly observed in the fault diagnosis domain. Next, employ the text mining algorithms that make use of this ontology to identify the necessary artifacts. This methodology of D-matrix construction consists of the following building blocks—document annotation, term extraction, and phrase merging. Following figure shows the flow of this method[12].

1. **PROBLEM DEFINATION**

Concept-based mining model consists of sentence-based concept analysis, document-based concept analysis, corpus-based concept-analysis, and concept-based similarity measure. The concept-based mining model can effectively discriminate between non-important terms with respect to sentence semantics and terms which hold the concepts that represent the sentence meaning. This method has drawback that it cannot work on web documents [1]. Ontology-based text mining method (OTMM) to cluster research proposals was a hybrid method for grouping research proposals for project selection. It uses text-mining, multilingual ontology, optimization, and statistical analysis techniques to cluster research proposals based on their similarities. The proposed OTMM was used together with statistical method and optimization models. Future work is needed to cluster external reviewers based on their research areas and to assign grouped research proposals to reviewers systematically [6].

Paper presentation slide generation (PPSGen) generates presentation slides from academic papers which trains a sentence scoring model based on support vector regression (SVR) and use the integer linear programming (ILP) method to align and extract key phrases and sentences for generating the slides. The advantage of this method is that it generates the slides automatically and it is well structured slides that can be easily understandable by humans. Additional information such as other relevant papers and the citation information can be used to improve the generated slides. This issue can be considered in the future [15]. Fuzzy self-constructing feature clustering (FFC) algorithm in which an incremental clustering approach to reduce the dimensionality of the features in text classification. Words that are similar to each other are grouped into the same cluster. FFC runs much faster than any other clustering method and also provide comparably good or better extracted features for classification. Future research in this can be done by considering the clustering method for faster recognition of required words[2].

1. **DESCRIPTION OF PRAPOSED WORK**
   1. **Proposed System**

Mainly system are divided into two modules. They are as follows

1. **User Module**
2. **Admin Module**

Above module contain the following activity.

1. **Research Base Paper as a input Collection from User**

Research paper selection is crucial and influential task for government and private research funding agencies. In these funding agencies large numbers of research proposals or papers are received, so that it common to group those proposals in order to groups form to their similarities in research disciplines. Funding agencies like NSFC maintain a directory of discipline areas that form a tree structure. As domain ontology, research ontology is a public concept set of the research project management domain. The research topics of different disciplines can be clearly expressed by research ontology.

1. **The processing Research Base Paper With all Keyword Found**

Each concept can analyze at the sentence level, a new concept-based frequency measure, called the frequency conceptual term is proposed. Sentence importance assessment is one of the two key steps, which aims to assign an importance score to each sentence in the given paper. The score of each sentence will be used at summarization process. In this study, we introduce a few useful features and High utility item sets .Generate group based on keyword search.

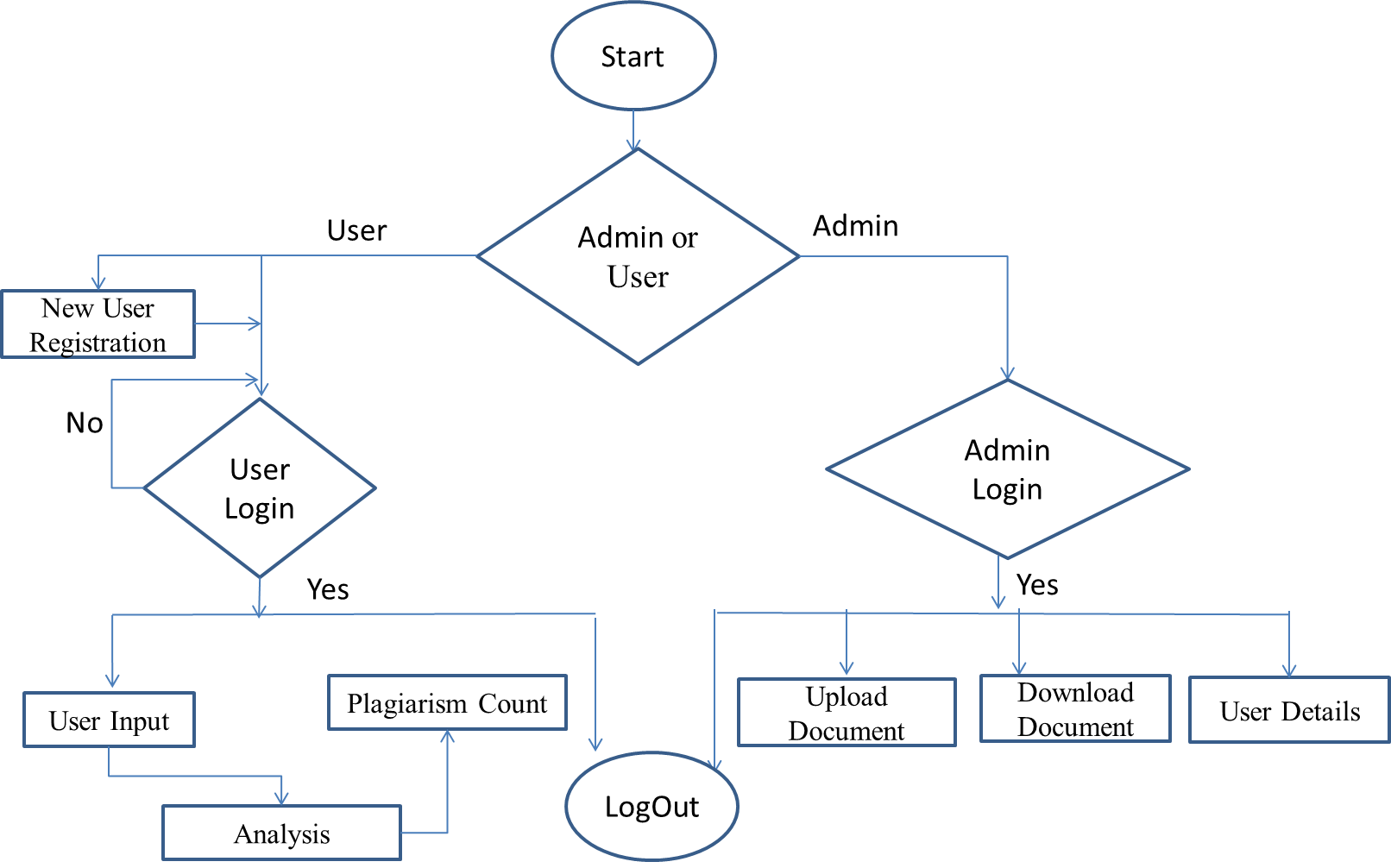
1. **The research keyword data are summarized**

After the research proposals are classified by the discipline areas, the proposals in each discipline are dividing into item sets using the text-mining technique. This step uses an Apriori algorithm to sets the feature vectors based on similarities of research areas. The Apriori algorithm is a typical unsupervised learning neural network model that item sets input data with similarities. The Apriori Algorithm is an influential algorithm it is used for mining frequent item sets for Boolean association rules. Apriori uses a "bottom up" approach, where one item at a time the frequent subsets are extended. It can form high utility item set if that item set is large then regrouping that group with closed utility item set using cluster. Finally generate summary of research document using generated cluster in each research area. These concept used for plagiarism analysis which can useful for calculate the percentage how much paper is copy paste. The graph based analysis can be obtained in it. The apriori algorithm used for text summarization and plagiarism concept.

1. **Find the Analysis of plagiaristic count**

Finally generate summary a count of research document using generated cluster in each research area. These concept can be used for plagiarism analysis which can useful for calculate the percentage how much paper is copy paste. The graph based analysis can be obtained in it.

**Flowchart :**

****

1. **IMPLECATION**

We have presented constructing feature clustering to generate summary with minimum time complexity, which is an incremental clustering approach to reduce the dimensionality of the features in text classification. Features that are similar to each other are grouped into the same cluster. Each cluster is characterized by a membership function with statistical mean and deviation. If a word is not similar to any existing cluster, a new cluster is created for this word. Similarity between a word and a cluster is defined by considering both the mean and the variance of the cluster. Experiments on three real-world data sets have demonstrated that our method can run faster and obtain better extracted features than other methods.

**References :**

[1] Shady Shehata, Fakhri Karray and Mohamed S. Kamel, “An Efficient Concept-Based Mining Model for Enhancing Text Clustering,” *IEEE Transaction on Knowledge And Data Engineering, VOL. 22, NO. 10,page no.1360-1371, OCTOBER 2010.*

[2] Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee,“ A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification”, *IEEE Transaction on Knowledge And Data Engineering, VOL. 23, NO. 3, page no.335-349, MARCH 2011.*

[3] Anindya Ghose and Panagiotis G. Ipeirotis, “Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics,” *IEEE Transaction on Knowledge And Data Engineering, VOL. 23, NO. 10] , page no1498-1512., OCTOBER 2011.*

[4] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu,“Effective Pattern Discovery for Text Mining,” *IEEE Transaction on Knowledge And Data Engineering, VOL. 24, NO. 1, page no.30-44, JANUARY 2012*.

[5] Xiang Wang, Xiaoming Jin, Meng-En Chen, Kai Zhang, and Dou Shen, “Topic Mining over Asynchronous Text Sequences,” *IEEE Transaction on Knowledge And Data Engineering, VOL. 24, NO. 1, page no.156-169, JANUARY 2012*