iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** IV **Month of publication:** April 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Searching Relevant Documents from Large Volume of Unstructured Database

Sarika Kolhe[1], Varsha Tambe[2], Gayatri Pawar[3], Priyanka Ubale[4] , Prof. Nihar Ranjan5
[1,2,3,4]*Computer Engineering, Sinhgad Institute Of Technology & Science, Pune,India.*
[5]*Asst.Prof. Computer Department, Sinhgad Institute Of Technology & science, Pune, India*

Abstract—In large organizations managing of data is very tedious task. these includes unstructured data such as images,videos,MP3 files, emails etc. The central aspect of research is to identify right document from unstructured documents. It refers Tf-Idf technology, clustering mechanism, similarity measure etc. When multiple document contains same data as input then document which is most similar to input query it should be display first. For that we can use Stemming Algorithm.
Index Terms—Text Mining, Classification, Unstructured Data, Clustering, Indexing, Information Retrieval, Query Formulation.

I. INTRODUCTION

Unstructured data can be categorized in two parts: Textual and Non-Textual. Unstructured data is in the form of Images, MP3 files, Videos can be classified as non textual unstructured data. Messages, memos ,Email, Word documents can be classified as textual unstructured data. Proposed system focuses on finding right documents from unstructured documents. For example Hiring the right candidate for the appropriate position in the organization. To select the perfect match from pool of applications. Organization creates job description which contains all requirements. Application which are relevant will be short listed and then exact document will be selected. For finding such kind of relevant documents we can use information retrieval system which satisfies users information needs. The main goal of Information retrieval is to find information relevant to users input within a collection of documents and which are relevant to users query.

II. ABOUT TEXT MINING

The text mining is subset of data mining. The process of deriving high quality information from text is called text mining. The purpose of text mining is to process unstructured data. It involves parsing, classification, retrieval, Indexing, clustering. Machine learning methodology is used in text mining. Text mining focus on association identifying trends in large collection of documents. Machine learning algorithm extract meaningful numeric indices from the text and make access to the contained information. The primary objective of text mining is to discover patterns among huge collection of documents. It focuses on identification and extraction of natural documents.

III. OVERVIEW OF THE SYSTEM

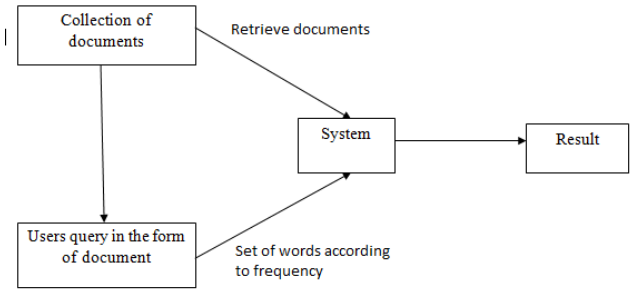
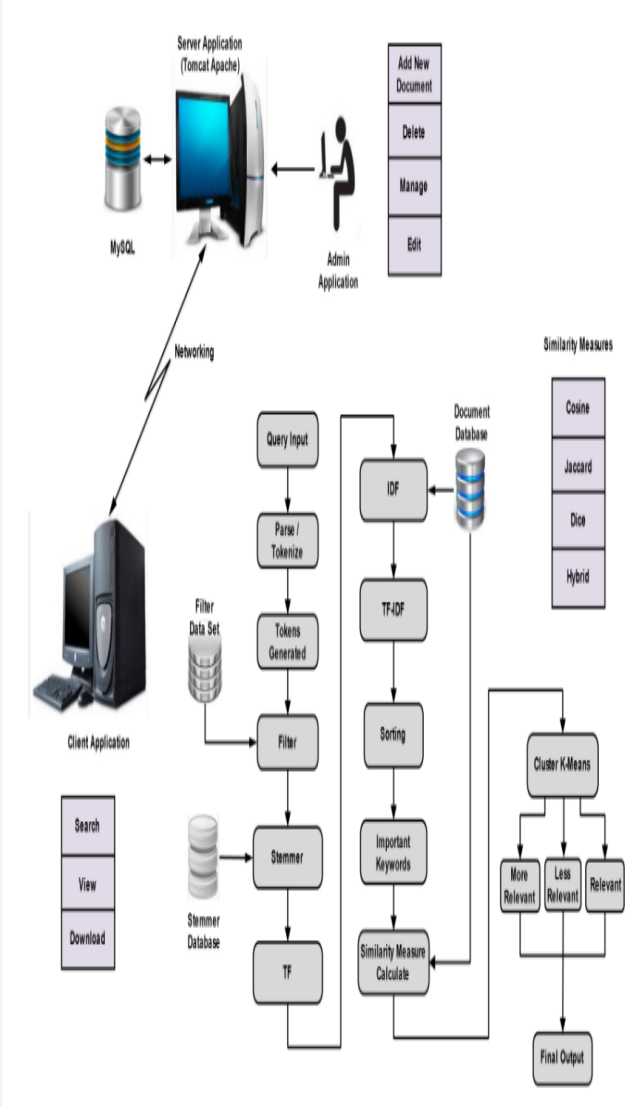


Fig.1 Overview of the system

IV. SYSTEM ARCHITECTURE



V. ISSUES HANDLED BY THE SYSTEM

A. Document Indexing

Indexing is the method which converts an unstructured collection of tokens which is in the form of words and tokens to a data structure called index. An index can directly point to a content of a collection of documents and of each document.
The keywords, index terms that is required to match relevant documents.

B. Stop Word List

Stop word list is used to prevent unwanted terms from been indexed. For example short term function words ‘the’, ‘is’, ‘at’, ‘which’, ‘and’, ‘on’ etc.Stop words are words which are filtered out before or after processing of natural language data.Many of frequently used words in English are worthless in text mining. These words are called as stop words.
Why To Remove stop words?
Reduce indexing file size.
Stop words accounts 20-30% of total word count.
Improve efficiency.
Stop words are not useful for searching. Stop words always have large no.of hits.

C. Synonyms

The word which has same meaning called as synonyms. For example “Hard” and “Difficult” or “happy” and “Cheerful”. In above example there are two different word but has same meaning.

D. Stemming Words

Stemming of word is the important step before indexing of input documents. It reduces the words to there roots. The words which has different grammatical forms are identified and called as same word. For example stemming will consider “work ”,worked” and “working” will be recognized by text mining program as same word.

1) Usefulness Of Stemming Words: Improving effectiveness of text mining & information retrieval. & Matching similar words. Reducing indexing size. Combing words with same roots may reduce indexing size as much as 40-50%.

2) Basic Stemming Method: Remove ending & Transform words

E. Classification

Text classification is process of categorize documents into predefined categorize according to their contents. Classification is basic step of text retrieval system which retrieve text in response to user query. For example classification of class student according to their percentage. Percentage above 60 will classify as first class and 55 above as higher second class etc.

VI. TECHNIQUES

For finding the relevant documents from unstructured documents we are going to use following techniques.

A. TF-IDF Technology

Term Frequency- Inverse Document Frequency

It is a weight frequently used in information retrieval and text mining. It is a statistical measurement tool which is used to evaluate how important a word is to a document in a collection of documents The Term frequency(TF) in the given document is the number of times a given term occurs in that document. this count is normalize to prevent bias towards longer documents to five a measure of important of the terms Ti within the particular documents Dj

Calculate the term frequency of the word. divide it by number of total words from the given documents

Calculate the inverse documents frequency (IDF). This is done by first dividing total number of documents by the number of documents that contain actual keyword in question. Then taking the logarithm of the result. Multiply the TF by the IDF, to get the result. For example: Lets calculate TF-IDF for the word like we counted 4 instances of the word like in the link building blog post. the number of total words in that blog post is 725. Also, 4 of the 7 blogs posts contains the word ‘like’ that gives us following calculations:

4/725=0.005517
Log(7/4)=0.2430
0.005517*0.2430=0.001341

B. Sorting Algorithm

Documents are classified based on discipline areas to which they belong. Simple sorting algorithm is used for classification. Suppose there are N discipline areas, Tk denotes area n(n=1.....N). Ki denotes documents i(i=1,2.....I), and Dk represents set of documents which belongs to area n. Then sorting algorithm can be implemented as follows.

TABLE I
Sorting Algorithm

```
For n = 1 to N
        For i = 1 to I
                If Ti belongs to Dk, then
                        Ti is added to Dk
        End
End
```

C. Similarity Measure

Similarity measure represent similarity between two documents to query or one documents or one query. Types of similarity measure are as follows: cosine, Jaccard coefficient, Euclidean ,Pearson correlation coefficient. It is a function that assigns a numerical value between 0 & 1 to pair of objects .if value is zero then two documents are totally different and if it is same then they are exactly identical. accurate clustering is mainly based on similarity measure. Semantic measurement: It is part of similarity measure. Semantic level consists concept releted to term in syntactic level. WordNet is used to calculate ascertain correlation among parts of speech- noun, verb ,adjective, adverb. The minimum unit (synset),represent an exact meaning of word. Further it classify into word, clarification & synonyms.

D. Clustering

It is a machine learning techniques. Definition of clustering: collection of similar kind of data objects called as cluster. And the process of partitioning set of data in a set of meaningful subclass is called clustering.

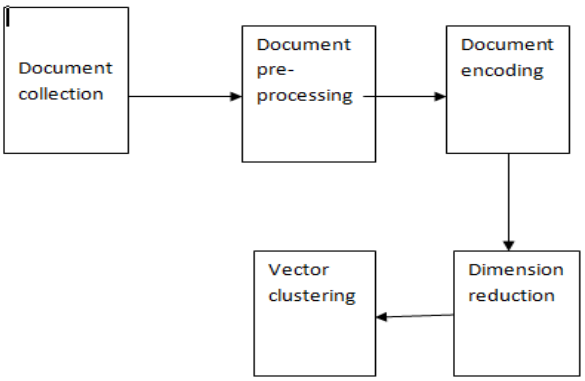


Fig. 2. main process of text mining.

E. K –Means Clustering

There are various methods of clustering. K-means is one of the most efficient methods for clustering. From the given set of n data, k different clusters; each cluster characterized with a unique centroid (mean) is partitioned using the K-means algorithm. The elements belonging to one cluster are close to the centroid of that particular cluster and dissimilar to the elements belonging to the other cluster.

Algorithm

Input: k: the number of clusters,

Output:

A set of k clusters.

Method:

Step 1: Choose k numbers of clusters to be determined.

Step 2: Choose Ck centroids randomly as the initial centers of the clusters.

Step 3: Repeat

3.1: Assign each object to their closest cluster center using Euclidean distance.

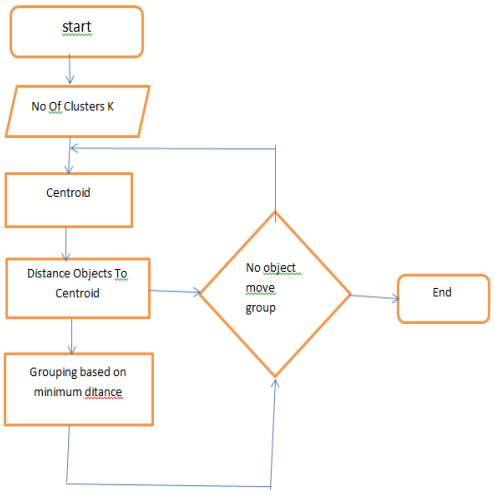
3.2: Compute new cluster center by calculating mean points.

Step 4: Until

4.1: No change in cluster center OR

4.2: No object changes its clusters.

Flowchart



Example of clustering

Indivisual	Variable1	Variable2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Step 1:

Initialization: Randomly we choose following two centroids (k=2) for two clusters.
In this case the 2 centroid are: m1=(1.0,1.0) and m2=(5.0,7.0).

Indivisual	Variable1	Variable2
1	**1.0**	**1.0**
2	1.5	2.0
3	3.0	4.0
4	**5.0**	**7.0**
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Step 2:

Thus, we obtain two clusters containing:

{1,2,3} and {4,5,6,7}.

Their new centroids are:

Indivisual	Centroid1	Centroid2
1	**0**	7.21
2(1.5,2.0)	**1.12**	6.10
3	**3.61**	3.61
4	7.21	**0**
5	4.72	**2.5**
6	5.31	**2.06**
7	4.30	**2.92**

Step 3:

Now using these centroids we compute the Euclidean distance of each object, as shown in table.

Therefore, the new clusters are:

{1,2} and {**3**,4,5,6,7}

Next centroids are: m1=(1.25,1.5) and m2 = (3.9,5.1)

Indivisual	Centroid1	Centroid2
1	**1.57**	5.38
2	**0.47**	4.28
3	2.04	**1.78**
4	5.64	**1.84**
5	3.15	**0.73**
6	3.78	**0.54**
7	2.74	**1.08**

Step 4:

The clusters obtained are:

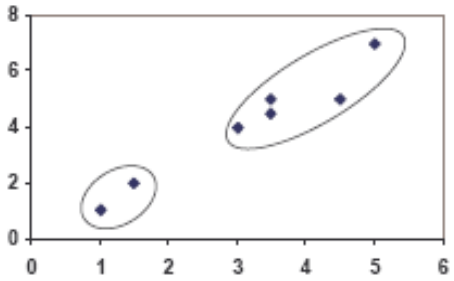
{1,2} and {3,4,5,6,7}

Therefore, there is no change in the cluster.

Thus, the algorithm comes to a halt here and final result consist of 2 clusters {1,2} and{3,4,5,6,7}.

Indivisual	Centroid1	Centroid2
1	**0.56**	5.02
2	**0.56**	3.92
3	3.05	**1.42**
4	6.66	**2.20**
5	4.16	**0.41**
6	4.78	**0.61**
7	3.75	**0.72**

PLOT:



VII. EVALUATION PARAMETERS

To validate proposed system several evaluation parameters are used.

A. Precision

The precision is a positive predictive value. It is fraction of retrieved instances that are relevant to the documents. Precision measures how precise a search is. the higher the precision, the less unwanted documents.

$$\text{Precision} = \frac{\text{No. of relevant documents retrieved}}{\text{Total no of documents retrieved}}$$

B. Recall

It is a fraction of relevant instances that are retrieve.Recall measures how complete a search is. the higher the recall, the less missing documents.

$$\text{Recall} = \frac{\text{No. of relevant documents retrieved}}{\text{No. of all relevant documents in database}}$$

C..False Positive And False Negative

In a classification task, the precision for a class is the number of true positives (i.e. the number of items correctly labeled as belonging to the positive class) divided by the total number of elements labeled as belonging to the positive class (i.e. the sum of true positives and false positives, which are items incorrectly labeled as belonging to the class). Recall in this context is defined as the number of true positives divided by the total number of elements that actually belong to the positive class (i.e. the sum of true positives and false negatives, which are items which were not labeled as belonging to the positive class but should have been). In information retrieval, a perfect precision score of 1.0 means that every result retrieved by a search was relevant (but says nothing about whether all relevant documents were retrieved) whereas a perfect recall score of 1.0 means that all relevant documents were retrieved by the search (but says nothing about how many irrelevant documents were also retrieved).

VIII. ADVANTAGES

Fast bifurcation into relevant and ir-relevantdocuments.
Customizable clustering
For sorting unstructured documents quickly.

IX. DISADVANTAGE

Requires large data sets for proper bifurcation.

X. APPLICATIONS

Clustering research papers - good for research publications
Clustering documents for forensics.
Scanning emails at email-server (bulk clustering required)

XI. CONCLUSION & FUTURE SCOPE

Proposed application extract words from document formats of pdf, doc. Term extractors can be introduced to extract terms from document format such as docx, html etc. It would be very nice to rate the terms from where they are finding documents. Higher rating should given to terms occur in heading then content and so on. We will continue to follow the technologies of this area and try to get more precise results.

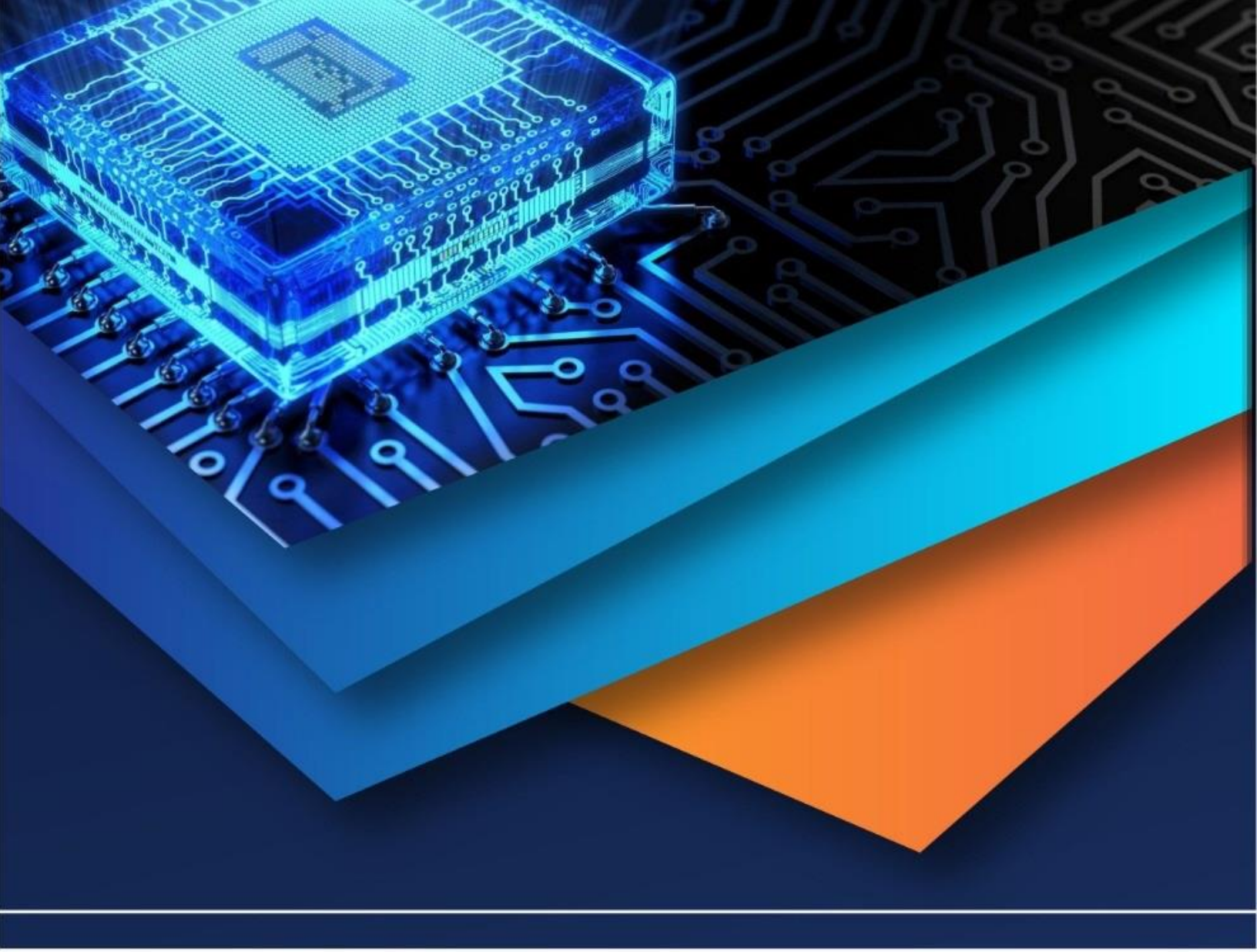
REFERENCES

[1] S. Siva Sathya, Philomina Simon, "Genetic Algorithm For Information Retrieval", Department of computer science.

[2]. Savidu Amarakoon, Amitha Caldera , Text mining:"finding right documents from large collection of unstructured documents. school of computing, university of Colombo.

[3]. Xueping Peng,Yujuan Cao," Mining web access log for the personalization recommendation", The school of computer science & technology,Beijing,China.

[4]. Jian Ma,Yong-Hong-Sun," An otology based text mining method to cluster proposals for research project selection".



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY