A PROJECT REPORT

ON

A Similarity Measure For Text Classification And Clustering

Submitted to the University of Pune, in the partial fulfillment of the requirements for the award of degree

of

BACHELOR OF COMPUTER ENGINEERING

BY

Rahul N Nalawade Exam No :

Akash G Samal Exam No :

DEPARTMENT OF COMPUTER ENGINEERING

STES'

SINHGAD ACADEMY OF ENGINEERING
S.No-40/4A, Kondhwa (Bk), Saswad Road,
Pune - 411048

YEAR 2015-16

SINHGAD TECHNICAL EDUCATION SOCIETY-PUNE

तमसो मा ज्योतिर्गमय

Sinhgad Institutes

SINHGAD ACADEMY OF ENGINEERING

S.No-40/4A, Kondhwa (Bk), Saswad Road,
Pune - 411048

CERTIFICATE

This is to certify that the project report entitled

“A Similarity Measure For Text Classification And Clustering”

Submitted by

Rahul N Nalawade	Exam No :
Akash G Samal	Exam No :

is a bonafide work carried out and is approved for the partial fulfillment of the requirement of
University of Pune, Pune for the award of the
Degree of Bachelor of Computer Engineering

This project work has not been earlier submitted to any other Institute or University for the award of any degree or diploma.

Place : Pune

Date :

Prof. ……	Prof.	
Internal Guide	Head,	Principal
Dept. of Computer Engineering	Dept. of Computer Engineering	
SAE, Pune	SAE, Pune	SAE, Pune

ACKNOWLEDGEMENT

It is an incidence of great pleasure in submitting this project report. Making this project reality takes many dedicated people and it is great pleasure to acknowledge the contribution of entire computer department.

We take this opportunity to express profound gratitude and ineptness for the personal involvement and constructive criticism provided beyond technical guidance during project to our Guide and project coordinator Prof.**Kiran Avhad**. of computer department. We shall ever be grateful to her for encouragement and suggestions given by her from time to time.

We should like to thank **H.O.D Prof. B.B.Gite** of Computer Department for providing the necessary facilities during the period of working of this project.

We should like to thank **Principal Prof. V.Wadhai** for providing the necessary facilities during the period of working of this project.

ThankYou,

Rahul N Nalawade

Akash G Samal

ABSTRACT

Measuring the similarity between documents is an important operation in the text processing field. In this project, a new similarity measure is proposed. To compute the similarity between two documents with respect to a feature, the proposed measure takes the following three cases into account: a) The feature appears in both documents, b) the feature appears in only one document, and c) the feature appears in none of the documents. For the first case, the similarity increases as the difference between the two involved feature values decreases. Furthermore, the contribution of the difference is normally scaled. For the second case, a fixed value is contributed to the similarity. For the last case, the feature has no contribution to the similarity. The proposed measure is extended to gauge the similarity between two sets of documents. The effectiveness of our measure is evaluated on several real-world data sets for text classification and clustering problems. The results show that the performance obtained by the proposed measure is better than that achieved by other measures.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 PROBLEM DEFINITION

Increase in the number of electronic documents it is hard to visualize these documents efficiently by putting manual effort. These have brought challenges for the effective and efficient organization of web page documents automatically. A similarity measure for text classification and clustering extracts features from web pages. On the basis of extracted features similarity between web pages are going to be calculated. Text document clustering methods attempt to segregate the documents into groups where each group represents some topic that is different than those topics represented by the other groups.

2 LITERATURE SURVEY

Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee proposed a new measure for computing the similarity between two documents. Several characteristics are embedded in this measure. It is a symmetric measure. The difference between presence and absence of a feature is considered more essential than the difference between the values associated with a present feature. The similarity decreases when the number of presence-absence features increases. An absent feature has no contribution to the similarity. The similarity increases as the difference between the two values associated with a present feature decreases. This work mainly focuses on textural features. Furthermore, the contribution of the difference is normally scaled. To improve the efficiency, they have provided an approximation to reduce the complexity involved in the computation. The results have shown that the performance obtained by the proposed measure is better than that achieved by other measures. Gaddam Saidi Reddy and Dr.R.V.Krishnaiah approach in finding similarity between documents or objects while performing clustering is multi-view based similarity. All measures such as cosine, Euclidean, Jaccard, and Pearson correlation are compared. The conclusion made here is that E cludean and Jaccard are best for web document clustering. They both selected related attributes for given subject and calculated distance between two values. Both of them used an algorithm known as Hierarchical Agglomerative Clustering in order to perform clustering. Their computational complexity is very high that is the drawback of these approaches. Proposed a similarity measure known as MVS (Multi-Viewpoint based Similarity), when it is compared with cosine similarity, MVS is more useful for finding the similarity of text documents. The empirical results and analysis revealed that the proposed scheme for similarity measure is efficient and it can be used in the real time applications in the text mining domain. It makes use of more than one point of reference as opposed to existing algorithms used for clustering text documents. Shady Shehata, Fakhri Karray and Mohamed S. Kamel mentioned that the most of the common techniques in text mining are based on the statistical analysis of a term, either word or phrase. Statistical analysis of a term frequency captures the importance of the term within a document only. Text mining model should indicate terms that capture the semantics of text. The mining model can capture terms that present the concepts of the sentence, which leads to discovery of the topic of the document. The mining model that analyzes terms on the sentence, document, and corpus levels are introduced, can effectively discriminate between non important terms with respect to sentence semantics and terms. The term which contributes to the sentence semantics is analyzed on the sentence, document, and corpus levels rather than the traditional analysis of the document only. It is important to note that extracting the relations between verbs and their arguments in the same sentence has the potential for analyzing terms within a sentence. The information about who is doing what to whom clarifies the contribution of each term in a sentence to the meaning of the main topic of that sentence. It is shown that the standard deviation is improved by using the concept-based mining model. Anna Huang declared that before clustering, a similarity/distance measure must be determined. The measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. It is very difficult to conduct a systematic study comparing the impact of similarity metrics on cluster quality, because objectively evaluation g cluster quality is difficult in itself. The clusters, which are generated in an unsupervised way, are compared to the pre-defined category structure, which is normally created by human experts. This kind of evaluation assumes that the objective of clustering is to replicate human thinking, so a clustering solution is good if the clusters are consistent with the manually created categories. It is found that there is no measure that is universally best for all kinds of clustering problems. The performance of the cosine similarity, Jaccard correlation and Pearson’s coefficient are very close, and are significantly better than the Euclidean distance measure experimented with the web page documents. Hung Chim and Xiaotie Deng found that the phrase has been considered as a more informative feature term for improving the effectiveness of document clustering. They proposed a phrase -based document similarity to compute the pairwise similarities of documents based on the Suffix Tree Document (STD) model. By mapp ing each node in the suffix tree of STD model into a unique feature term in the Vector Space Document (VSD) model, the phrase -based document similarity naturally inherits the term tf-idf weighting scheme in computing the document similarity with phrases. They applied the phrase-based document similarity to the group-average Hierarchical Agglomerative Clustering (HAC) algorithm and developed a new document clustering approach. Their evaluation experiments indicate that the new clustering approach is very effective on clustering the documents of two standard document benchmark corpora OHSUMED and RCV1.Finally they found that both the traditional VSD model and STD model play important roles in text-based information retrieval. The concept of the suffix tree and the document similarity are quite simple, but the implementation is complicated. Investigation is required to improve the performance of the document similarity. They conclude that the feature vector of phrase terms in the STD model can be considered as an expanded feature vector of the traditional single-word terms in the VSD model.Yanhong Zhai and Bing Liu studied the problem of extracting data from a Web page that contains several structured data records. The objective is to segment these data records, extract data items/fields from them and put the data in a database table.They proposed approach to extract structured data from Web pages. Although the problem has been studied by several researchers, existing techniques are either inaccurate or make many strong assumptions. Inderjit Dhillon, Jacob Kogan & Charles Nicholas found that in particular, when the processing task is to partition a given document collection into clusters of similar documents a choice of good features along with good clustering algorithms is of paramount importance. Feature or term selection along with a number of clustering strategies. The selection techniques significantly reduce the dimension.Syed Masum Emran and Nong Ye said distance metric value is used to find the similarity or dissimilarity of the current observation from the already established normal profile. To find the distance between normal profile and current observation value, one can use many distance metrics. Alexander Strehl, Joydeep Ghosh, and Raymond Mooney studied if clusters are to be meaningful, the similarity measure should be invariant to transformations natural to the problem domain. The features have to be chosen carefully. They conducted a number of experiments to assure statistical significance of results. Metric distances such as Euclidean are not appropriate for high dimensional, sparse domains. Cosine, correlation and extended Jaccard measures are successful in capturing the similarities implicitly indicated by manual categorizations as they seen for example in Yahoo. S. Kullback and R. A. Leibler found that in terms of similarity measure for information retrieval, difficult it is to discriminate between the populations. R. A. Fisher introduced the criteria for sufficiency required that the statistic chosen should summarize the whole of the relevant information supplied by the sample. Mei-Ling Shyu, Shu-Ching Chen, Min Chen & Stuart H. Rubin mentioned that compared to the regular documents, the major distinguishing characteristics of the Web documents is the dynamic hyper structure. In their experimental results they found that the Euclidean distance gives the worst performance, followed by the cosine coefficient

Existing System

First, about the determination of K value. Through the analysis, the K value of the initial cluster centers to determine the far-reaching impact throughout the clustering process and the final clustering results, while the K value in practical applications is very difficult to direct or one-time determination . Especially, if the amount of data tends to infinity which is pending, the K value of the K-means algorithm to determine will be very difficult. At present, there are two clustering algorithms to determine the K value is relatively effective which is the cost function based on distance and propagation clustering algorithm based on nearest neighbors. The former find the minimum through using the cost function. Thus obtain the corresponding K value. The latter using nearest neighbor clustering algorithm to calculate the appropriate number of cluster center, the number of cluster center provides for the maximum K value of the K-means clustering algorithm to get the optimal value of K. Second, about the choice of initial cluster centers. K-means clustering algorithm using the iterative method to solve the problem, except the first step, the clustering results of each step are improved to some extent; otherwise terminate the process of iteration. Traditional K-means clustering algorithm takes the cluster squares error and the criterion function value change or not as the iterative termination conditions. But the clustering results obtained from this criterion function easily fall into local minimum solution, the result is the clustering results of search are moving toward the direction of diminishing the criterion function value.

Procedure of K-means Algorithm

1) Distribute all objects to K number of different cluster at random;

2) Calculate the mean value of each cluster, and use this mean value to represent the cluster;

3) Re-distribute the objects to the closest cluster according to its distance to the cluster center;

4) Update the mean value of the cluster. That is to say, calculate the mean value of the objects in each cluster;

5) Calculate the criterion function E, until the criterion function converges

$E = \sum k\ j=1 \sum n\ i=1\ xi \epsilon cj\ \|xi - mj\|\ 2$

6) In which, E is total square error of all the objects in the data cluster, xi bellows to data object set, mi is mean value of cluster Ci (x and m are both multi-dimensional). The function of this criterion is to make the generated cluster be as compacted and independent as possible.

Limitations:

Limitation 1: Handling Empty Clusters: One of the problems with the basic K-means algorithm given earlier is that empty clusters can be obtained if no points are allocated to a cluster during the assignment step. If this happens, then a strategy is needed to choose a replacement centroid, since otherwise, the squared error will be larger than necessary.

Limitation 2: Difficult to measure the no of clusters: The user has to choose the value of K, the number of clusters .Although for 2D data this choice can easily be made by visual inspection, it is not so for higher dimension data, and there are usually no clues as to what number of clusters might be appropriate.

3 SOFTWARE REQUIREMENTS SPECIFICATION

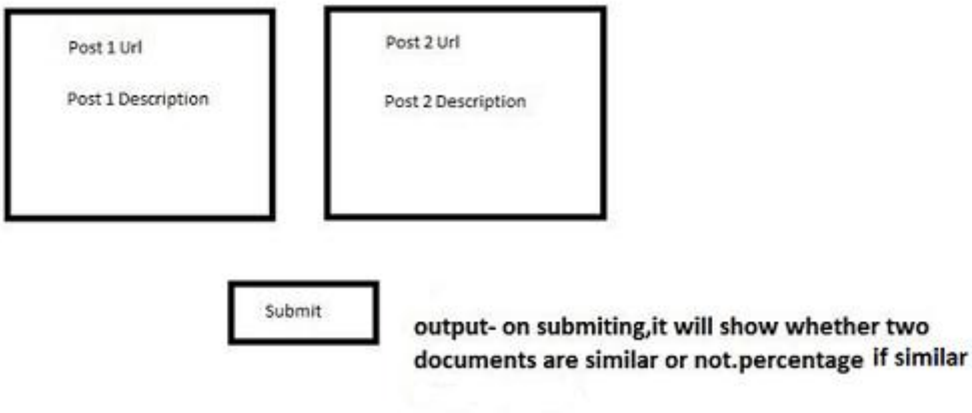
3.1 INTRODUCTION

The amazing progress of computer technology in the few decades has led to large supplies of powerful and affordable computers. Increase in the number of electronic documents it is hard to visualize these documents efficiently by putting manual effort. These have brought challenges for the effective and efficient organization of web page documents automatically. Extracting features from web pages is initial task found in mining. On the basis of extracted features similarity between web pages are going to be calculated. There is various similarity measures are pointed out for work. To implement the efficient similarity measure one has to do survey on outcomes.

Data mining is the process of extracting the implicit, previously unknown and potentially useful information from data. Document clustering, subset of data clustering, organizes documents into different groups called as clusters, where the documents in each cluster share some common properties according to defined similarity measure. Document clustering algorithms play an important role in helping users to effectively navigate, summarize and organize the information. Due to explosive growth of accessing information from the web, efficient access and exploration of information are needed critically. The Text processing plays an important role in information retrieval, data mining, and web search. Text mining attempts to discover new, previously unknown information by applying techniques from data mining. Clustering ,one of the traditional data mining techniques is an unsupervised learning paradigm where clustering methods try to identify inherent groupings of the text documents, so that a set of clusters is produced in which clusters exhibit high intra-cluster similarity and low inter-cluster similarity. Generally, text document clustering methods attempt to segregate the documents into groups where each group represents some topic that is different than those topics represented by the other groups.

The similarity measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. Before Clustering, a similarity/distance measure must be determined .Choosing an appropriate similarity measure is also crucial for cluster analysis ,especially for a particular type of clustering algorithms.

Text Categorization (TC) is the classification of documents with respect to a set of one or more pre-existing categories .The classification phase consists of generating weighted vector for all categories, then using a similarity measure to find the closest category. The similarity measure is used to determine the degree of resemblance between two vectors. To achieve reasonable classification results, a similarity measure should generally respond with larger values to documents that belong to the same class and with smaller values otherwise. During the last decades, a large number of methods proposed for text categorization were typically based on the classical Bag-of-Words model where each term or term stem is an independent feature.



output- on submiting,it will show whether two documents are similar or not.percentage if similar

Fig:1 Basic Structure of Similarity Measure

Two documents are taken for similarity measure. After submitting by clicking on submit button. It will show whether two documents are similar or not .If similar then display percentage of similarity .

3.1.1 Project Scope

In this project we will crawl (web scraping) data from different Bollywood sources. Once we get the data then, we will apply similarity algorithm.Here, we will be considering only bollywood article (news), once we get the news or article from different sources, we will be applying k-means algorithm on them to get the similar article. This will be unsupervised learning algorithm, by doing this we can calculate the similarity of any news website and same news is coming from different sources. The data we will be interested in Percentage, word match count and sources (eg.mid-day-ndtv-bollywoodhungama). Once we get the above data, we can show only one article and all other sources in bottom, so there will be no duplication of news.

3.1.2 User Classes and Characteristics

User:

- User supplies search query to search engine.Search Engine gives the result to clustering engine .After performing preprocessing similarity measure performs and clustering is done then result is display to user.The one article is at the top and others are at bottom.
- User can add new site to system.
- User view the status of newly added service.

3.1.3 Design and Implementation Constraints

Proposed System

Improved K-means Clustering Algorithm

Optimize the initial cluster centers, to find a set of data to reflect the characteristics of data distribution as the initial cluster centers, to support the division of the data to the greatest extent. Optimize the calculation of cluster centers and data points to the cluster center distance, and make it more match with the goal of clustering.In computing the region of high-density data set D, set $2\gamma = 1/100$, the entire area is divided into 100 parts, set the density threshold $\tau = n/100$.

Algorithm : Improved K-means Algorithm Input: data set x contains n data points; the number of cluster is k. Output: k clusters of meet the criterion function convergence. Program process:

Step 1. Initialize the cluster center.

Step 1.1. Select a data point xi from data set X, set the identified as statistics and compute the distance between xi and other data point in the data set X. If it meet the distance threshold, then identify the data points as statistics, the density value of the data point xi add 1.

Step 1.2. Select the data point which is not identified as statistics, set the identified as statistics and compute its density value. Repeat Step 1.2 until all the data points in the data set X have been identified as statistics.

Step 1.3. Select data point from data set which the density value is greater then the threshold and add it to the corresponding high-density area set D.

Step 1.4. Filter the data point from the corresponding high-density area set D that the density of data points relatively high, added it to the initial cluster center set. Followed to find the k-1 data points, making the distance among k initial cluster centers are the largest.

Step 2. Assigned the n data points from data set X to the closet cluster.

Step 3. Adjust each cluster center K .

Step 4. Calculate the distance of various data objects from each cluster center by formula and redistribute the n data points to corresponding cluster.

Step 5. Adjust each cluster center K .

Step 6. Calculate the criterion function E, to determine whether the convergence, if convergence, then continue; otherwise, jump to Step 4.

Advantages & Disadvantages:

K-Means Advantages:

1) If variables are huge, then K-Means most of the times computationally faster than hierarchical clustering, if we keep k smalls.
2) K-Means produce tighter clusters than hierarchical clustering , especially if the clusters are globular.
3) K-mean value algorithm is a classic algorithm to resolve cluster problems; this algorithm is relatively simple and fast.
4) For large data collection, this algorithm is relatively flexible and high efficient, because the Complexity is O (ntk). Among which, n is the times of iteration, k is the number of cluster, t is the times of iteration. Usually, k"n and t"n. The algorithm usually ends with local optimum..
5) It provides relatively good result for convex cluster.
6)Because the limitation of the Euclidean distance. It can only process the numerical value, with good geometrical and statistic meaning.

K-Means Disadvantages

1) The K value is most important for K-means clustering algorithm. There is no applicable evidence for the decision of the value of K (number of cluster to generate), and sensitive to initial value, for different initial value, there may be different clusters generated.

2) K-means clustering algorithm has a higher dependence of the initial cluster centers. If the initial cluster center is completely away from the cluster center of the data itself, the number of iterations tends to infinity, but also makes it easier for the final clustering results into local optimization, resulting in incorrect clustering results.

3) K-means clustering algorithm has a strong sensitivity to the noise data objects. If there is a certain amount of noise data in dataset, it will affect the final clustering results, leading to its error.

4) K-means clustering algorithm for the discovery of clusters of arbitrary shape is most difficult.

5) K-means clustering algorithm has main limitation on amount of data. In the iterative process, every time you need to adjust the cluster to which data object belongs and compute cluster center, so in case of large amount of data, the K-means clustering algorithm is not applicable.

To overcome this k-means disadvantages which algorithm we have to use

1)To overcome this we should have some initial boundary constraints on the k values. or we could use optimization algorithms to minimise the cross-validation error.

3.2 SYSTEM FEATURES

3.2.1 This will be unsupervised learning , by doing this we can calculate the similarity of any news website and same news is coming from different sources.make cluster of same articles and avoid duplication

3.2.2 we should have some initial boundary constraints on the k values. or we could use optimization algorithms to minimise the cross-validation error.

3.3 EXTERNAL INTERFACES

3.3.1 USER INTERFACES

Cluster Management

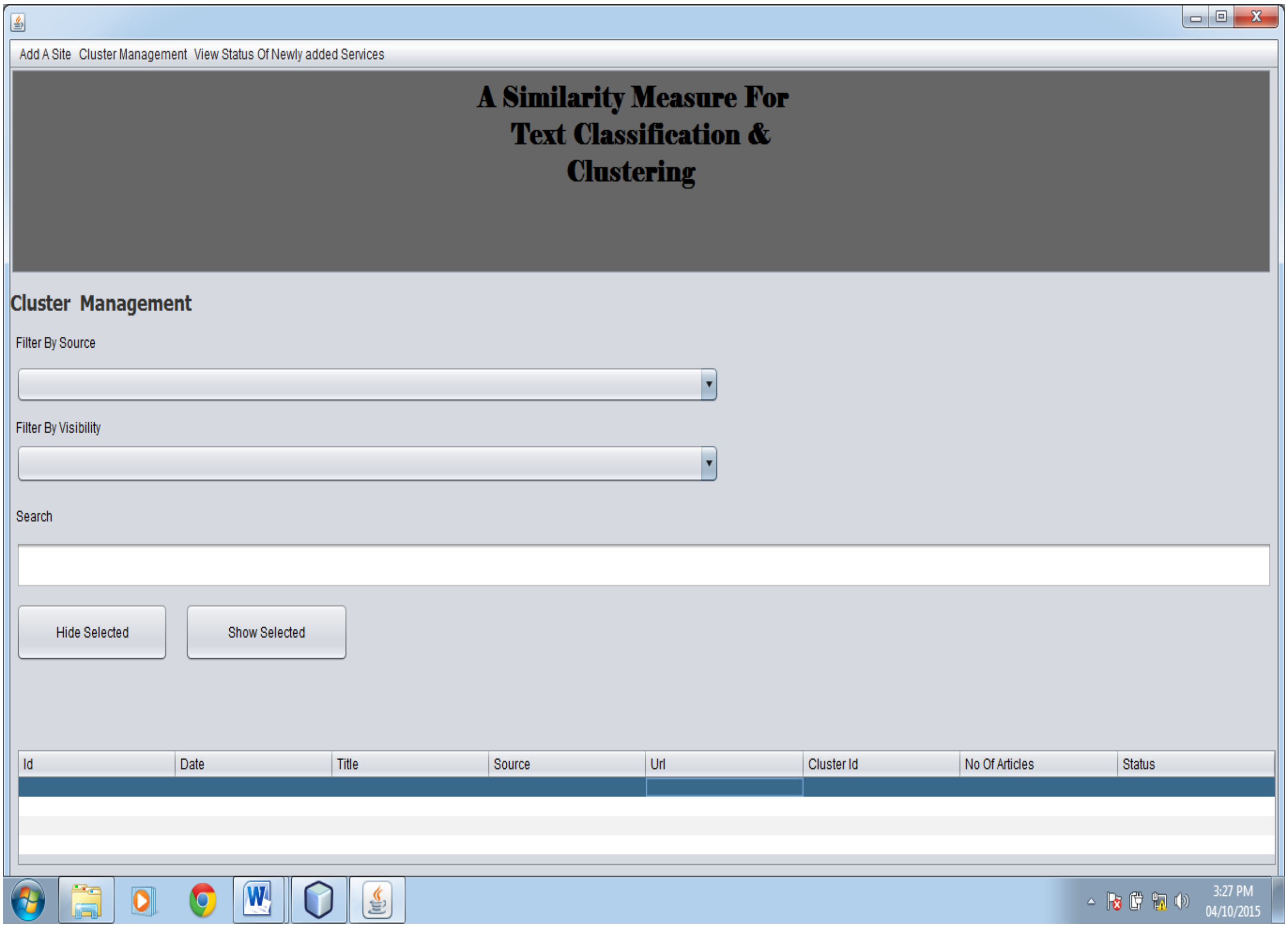
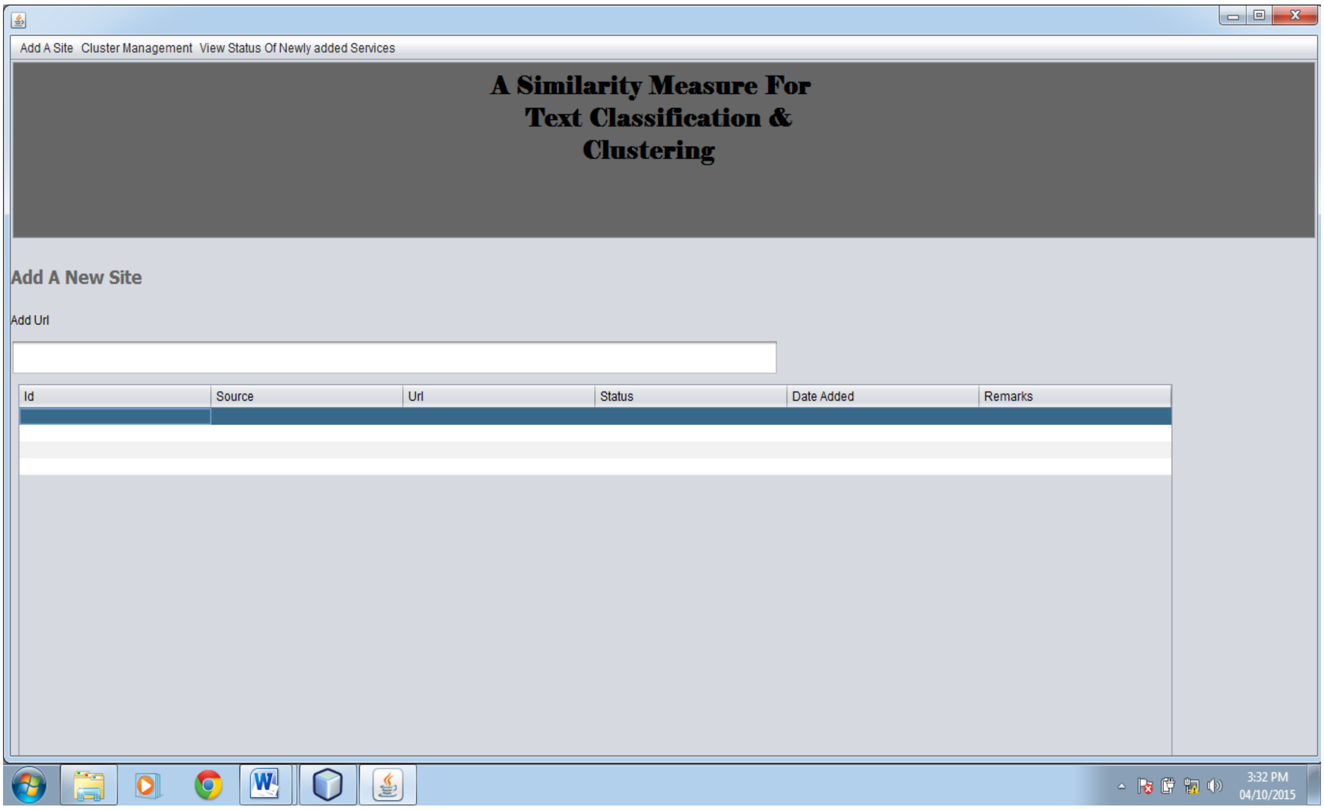


Fig:2 Cluster Management

Add A New Site



A Similarity Measure For Text Classification & Clustering

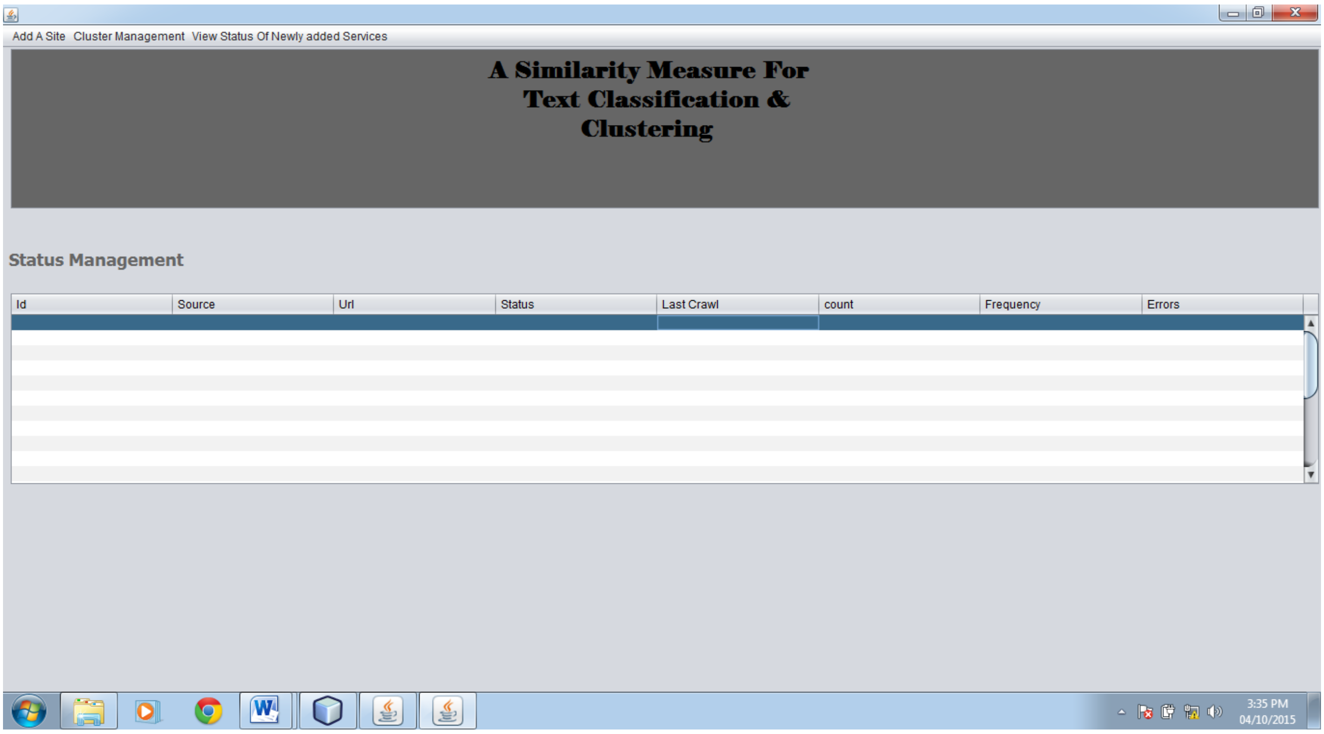
Add A New Site

Add Url

Id	Source	Url	Status	Date Added	Remarks

Fig:3 Add A New Site

Status Management



A Similarity Measure For Text Classification & Clustering

Status Management

Id	Source	Url	Status	Last Crawl	count	Frequency	Errors

Fig:4 Status Management

3.3.2 Hardware Requirements

- **General Hardware Requirements**

Component	Server Side
Architecture	32 Bit or 64 Bit
Processor	Pentium IV Processor Or above
RAM	512 mb RAM or more

Table :1 Hardware Requiremen

3.3.3 Software Requirement

- **General Software Requirements**

Component	Server Side
Software Required at Software Development Phase	• JDK 1.6.0 • Net bean 7.1 • Tomcat Apache Server • MongoDB
Operating System	Windows XP(or above) (32 or 64 Bits)

Table :2 Software Requirement

3.4 NON-FUNCTIONAL REQUIREMENTS

3.4.1 Performance Requirements

checking the fact that the system must perform as what every user expects .So in every action-response of the system, there are no immediate delays. In case of opening windows forms, of popping error messages and saving the settings or sessions there is delay much below 2 seconds ,In case of opening databases, sorting questions and computing there are no delays and the operation is performed in less than 2 seconds for opening, sorting, computing > 95% of the files.

3.4.2 Safety Requirements

Consistency: checking the fact that all users must be attachable to one server, so there would be appropriate control of the similarity statistics and information.Also in case of a potential loss of connection between the user and the server the user workprogress so far is lost. When the user finishes its work then its progress is sent to the server and be logged. In case of a potential server breakdown only the so far finished work is saved.

3.4.3 Software Quality Attributes

Availability: Checking that the system always has something to function and always pop up error messages in case of component failure. In that case the error messages appear when something goes wrong so to prevail availability problems.

Usability: Checking that the system is easy to handle and navigates in the most expected way with no delays. In that case the system program reacts accordingly and transverses quickly between its states.

Functionality: Checking that the systems provide the right tools for editing question Databases, creating clusters. In that case the tools that the Database editor provides are the ones that provide that attribute.

3.5 ANALYSIS MODELS

3.5.1 Data Flow Diagrams

Level 0 DFD

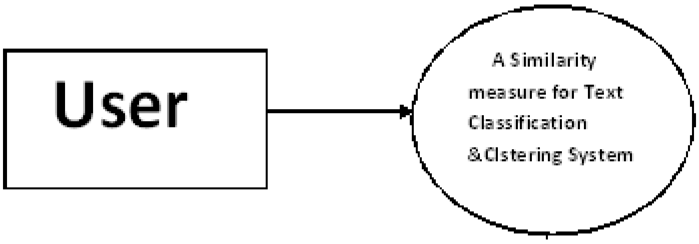

User
A Similarity measure for Text Classification &Clstering System

Fig:5 Level 0 DFD

Level 1 DFD

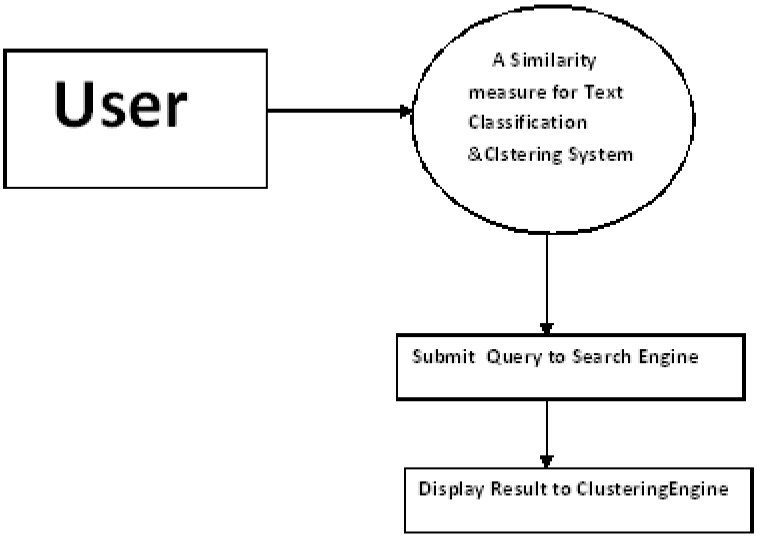

User
A Similarity measure for Text Classification &Clstering System
Submit Query to Search Engine
Display Result to ClusteringEngine

Fig:6 Level 1 DFD

Level 2 DFD

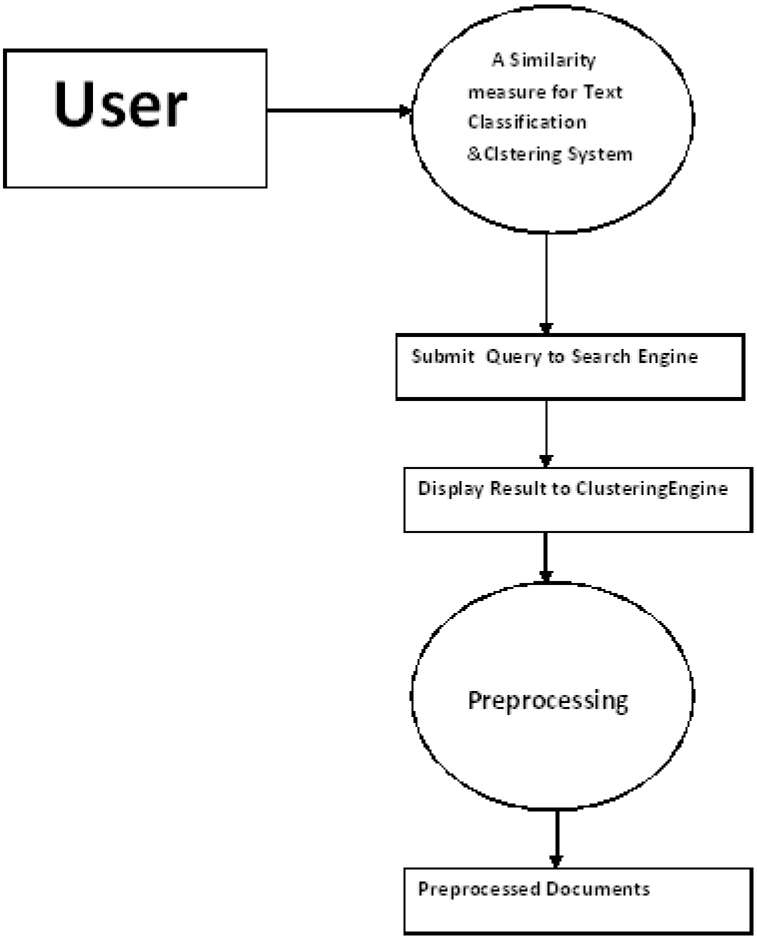

User
A Similarity measure for Text Classification &Clstering System
Submit Query to Search Engine
Display Result to ClusteringEngine
Preprocessing
Preprocessed Documents

Fig:7 Level 2 DFD

Level 3DFD

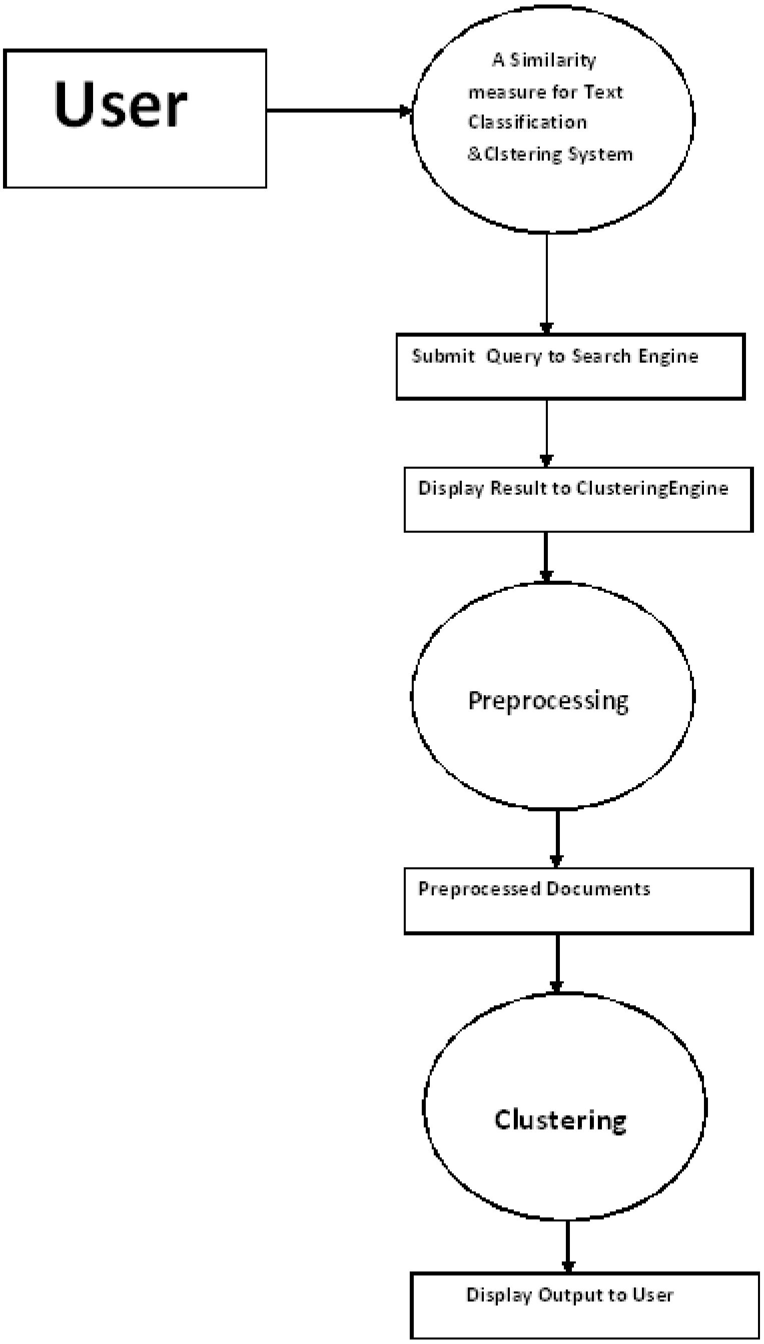

User
A Similarity measure for Text Classification &Clstering System
Submit Query to Search Engine
Display Result to ClusteringEngine
Preprocessing
Preprocessed Documents
Clustering
Display Output to User

Fig:8 Level 3 DFD

3.5.2 Entity relationship Diagrams

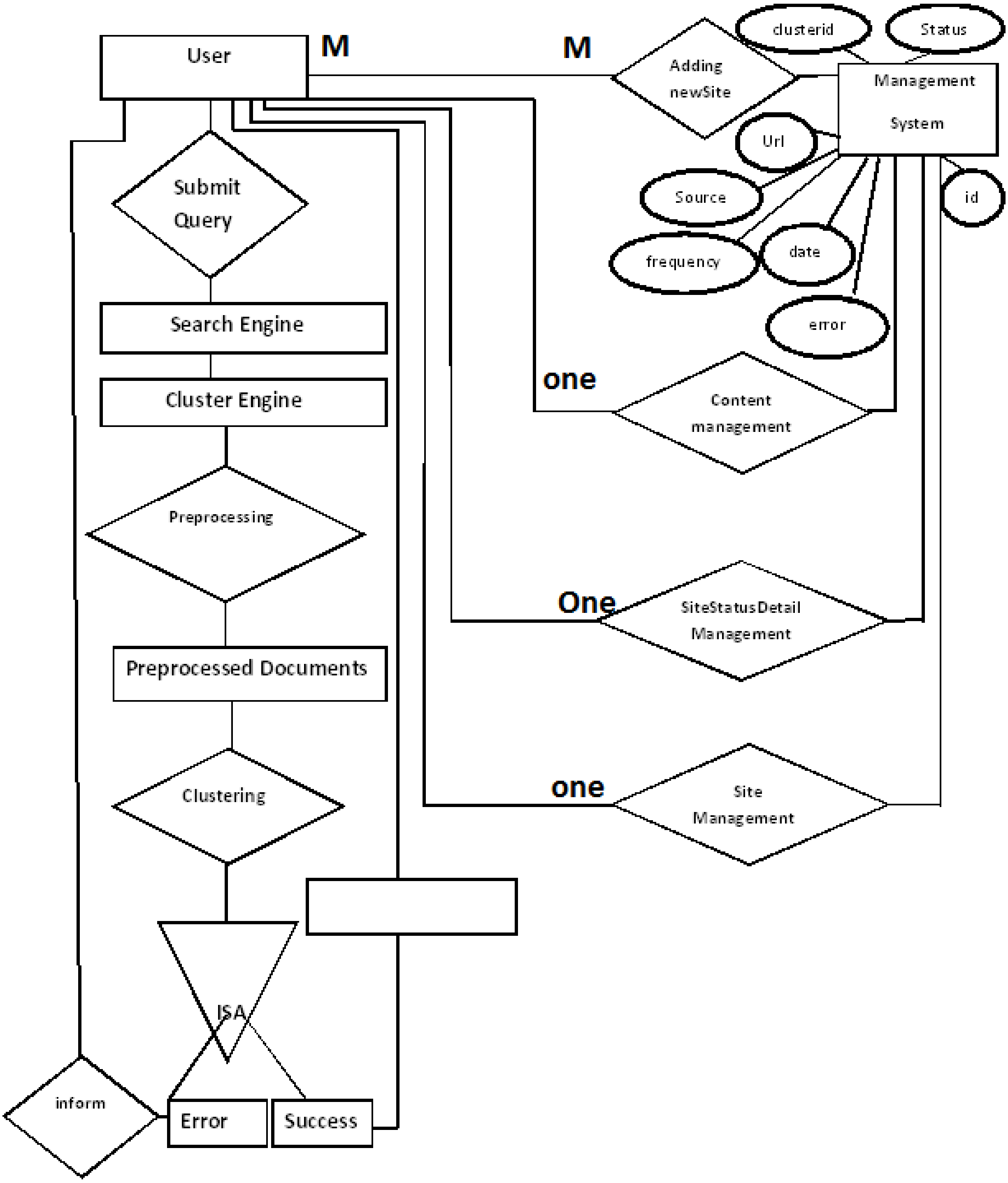


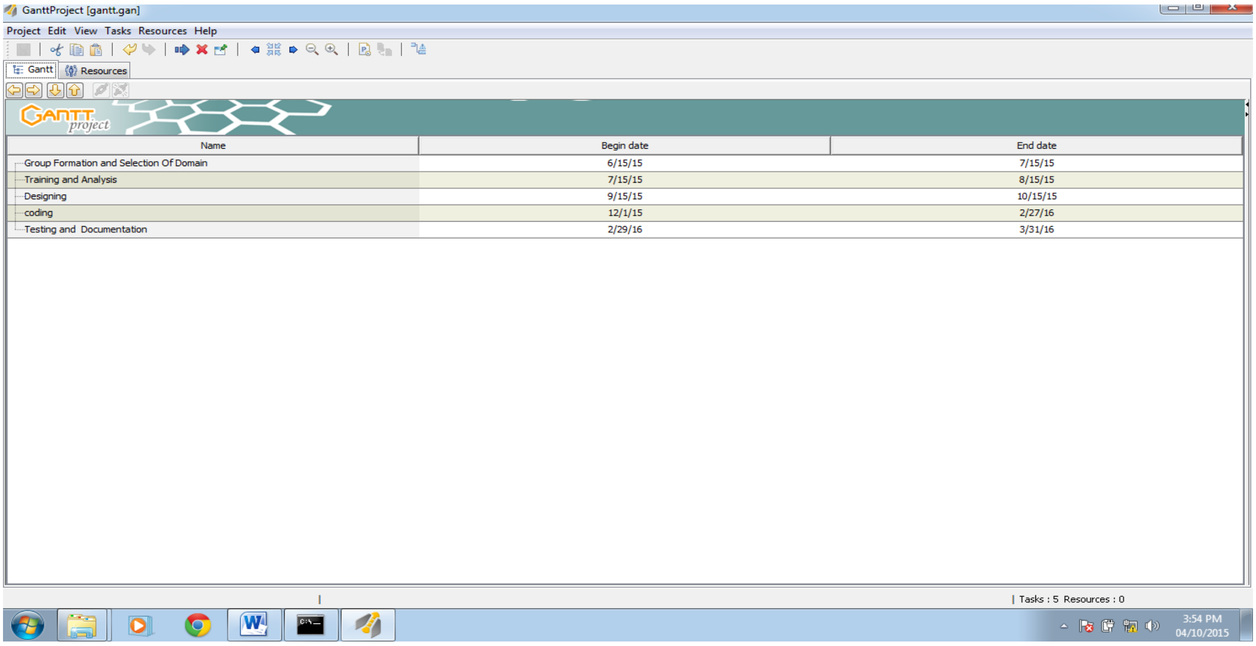
Fig:9Entity relationship Diagram

3.6 SYSTEM IMPLEMENTATION PLAN

Month	Target
June-July 2015	Group formation and selection of domain
Aug – Sept ‘15	Training
	Analysis
Oct ‘15	Designing
Dec ‘15	Coding
Mar ’16	Testing
	Documentation

Table:3 System Implementation Plan

Gantt Chart



Name	Begin date	End date
Group Formation and Selection Of Domain	6/15/15	7/15/15
Training and Analysis	7/15/15	8/15/15
Designing	9/15/15	10/15/15
coding	12/1/15	2/27/16
Testing and Documentation	2/29/16	3/31/16

Tasks : 5 Resources : 0

Table:4 Gantt Chart details

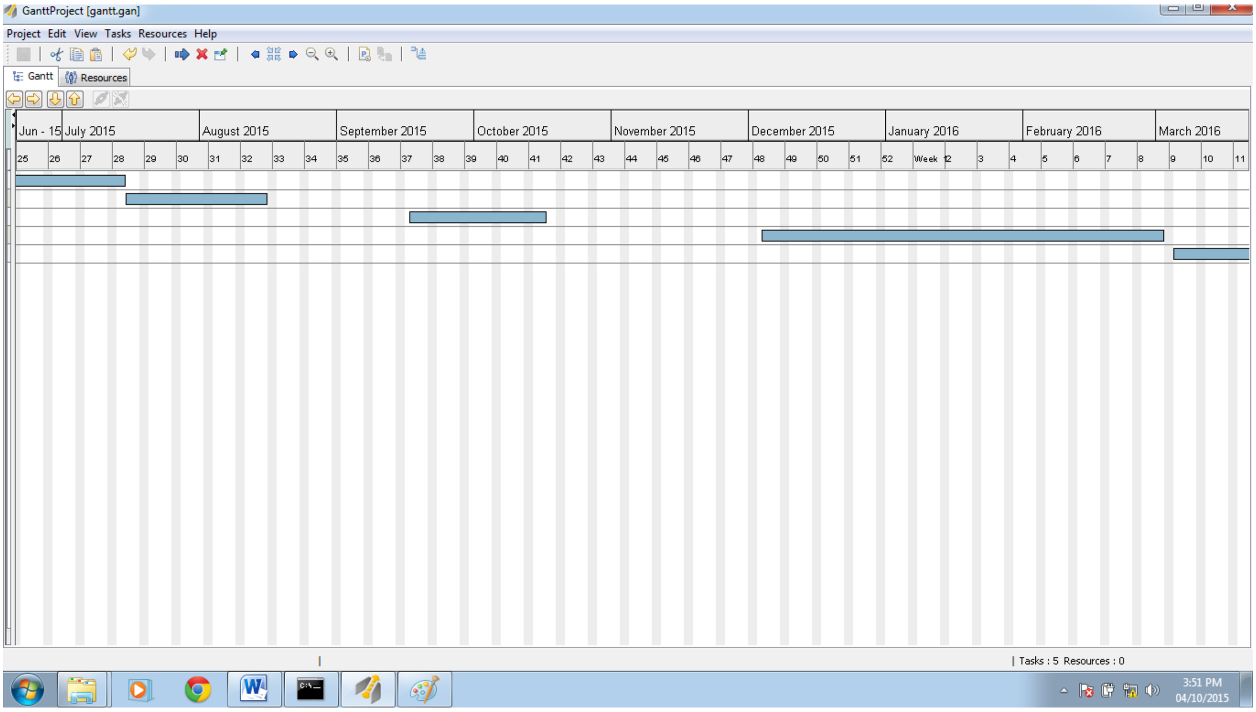


Fig:10 Gantt Chart

4 SYSTEM DESIGN

4.1 SYSTEM ARCHITECTURE

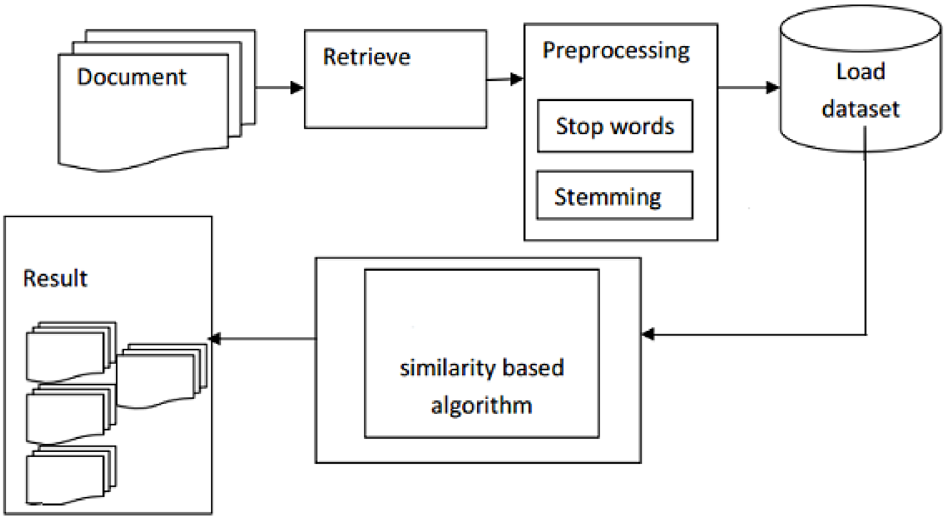


Fig:11 System Architecture

The Number of documents are retrived or crawl from different sources and Preprocessing is done on it. Following is the three steps of preprocessing

The results obtained by conventional search engine are preprocessed in-order to represent best binary vector space model representation oftext document. For that we have to follow some preprocessing steps.

These steps are

4.1.1 Stop words Removal

4.1.2 Tokenization

4.1.3 Stemming

4.1.1 Stop words removal

Many times, it makes sense to not index "stop words" during the indexing process. Stop words are words which have very little informational content. These are words such as: and, the, of, it, as, may, that, a, an, of, off, etc. Studies have shown that by removing stop words from the index, you may benefit with reduced index size without significantly affecting the accuracy of a user's query. Care must be taken however to take into account the user's needs. For example, the phrase "to be or not to be" from Hamlet is composed entirely of stop words. Most of the internet's search engines eliminate all the stop words from their indexes. By eliminating stop words from the index, the index size is typically reduced by about 33% for a word level index. For a record level index, then eliminating stop words is not typically done as they will not add significantly to the index size.After stop words removal tokenization and steeming is performed.

4.1.2 **Tokenization**

Tokenization is the process of replacing sensitive data with unique identification symbols that retain all the essential information without compromising its security.

4.1.3 **Stemming**

The concept stemming has been applied to information systems from their initial automation in 1916,s. the original goal of stemming was to improve performance and require less system resources by reducing the number of unique words that a system has to contain. Stemming algorithms are used to improve the efficiency of the information system and to improve recall. Stemmingis the process for reducing inflected (or sometimes derived) words to their stem, base or root form – generally a written word form. There are many stemming algorithms are available.

After preprocessing a similarity measure algorithm is applied and cluster is formed. Assign the documents to cluster and result is display to user

4.2 UML DIAGRAMS

The unified modeling language is a standard language for specifying, Visualizing, Constructing and documenting the software system and its components. It is a graphical language which provides a vocabulary and set ofsemantics and rules. The UML focuses on the conceptual and physical representation of the system. It captures the decisions and understandings about systems that must be constructed. It is used to understand, design, configure, maintain and control Information about the systems.

6.1.2 Aims of Modeling

The following are aims of modeling:

- Models help us to visualize a systemas it is or as we want to be
- Models permit us to specify the structure of system.
- Model gives us template that guides us in constructing system.
- Models can document the decisions we have made.

The vocabulary of the UML encompasses 3 kinds of building blocks.

They are

- Things
- Relationships
- Diagrams

There are 4 kinds of things in the UML.

- Structural things
- Behavioral things
- Grouping things
- Annotational things

There are 4 kinds of relationships

- Dependency
- Association
- Generalization
- Realization

Dependency:A dependency is a semantic relationship between 2 things in which a change to one thing affects the semantics of the other things.

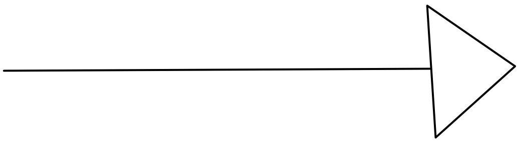
Symbol: —— —— —— —— ——▶

Association: An association is a structural relationship that describes a set of links, a link being a connection among objects.

Symbol: ————————————————

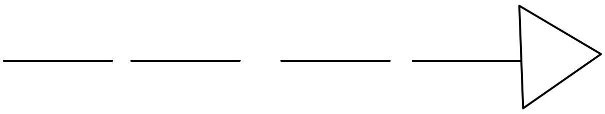
Generalization: A Generalization is a relationship in which objects of the specialized elements are suitable substitutable for objects of the generalized element. In Generalization, the relationship exist between a general thing and a kind of relationship.

Symbol:



Realization: A realization is a semantic relation between classifiers, where in one classifier guarantees to carry out and represented as dotted line with a hollow arrow head, pointing to the parent. It is a combination of dependency & Generalization.

Symbol:



Diagrams in the UML:

A diagram is the graphical presentation of a set of elements, mostoften rendered as a connected graph of vectors and arcs.

Diagrams in the UML :

Class Diagram

Activity Diagram

Sequence Diagram

Use-Case Diagram

Deployment Diagram

In our project we designed our system by using the following Diagrams.

4.2.1 Class Diagram

It shows a set of classes, interfaces and collaborations and their relationships. It addresses the static process, view of the system Classes involved in the above class diagram,

- User
- Search Engine
- Clustering Engine
- Preprocessing
- Stop-words Removal

- Tokenization
- Stemmin

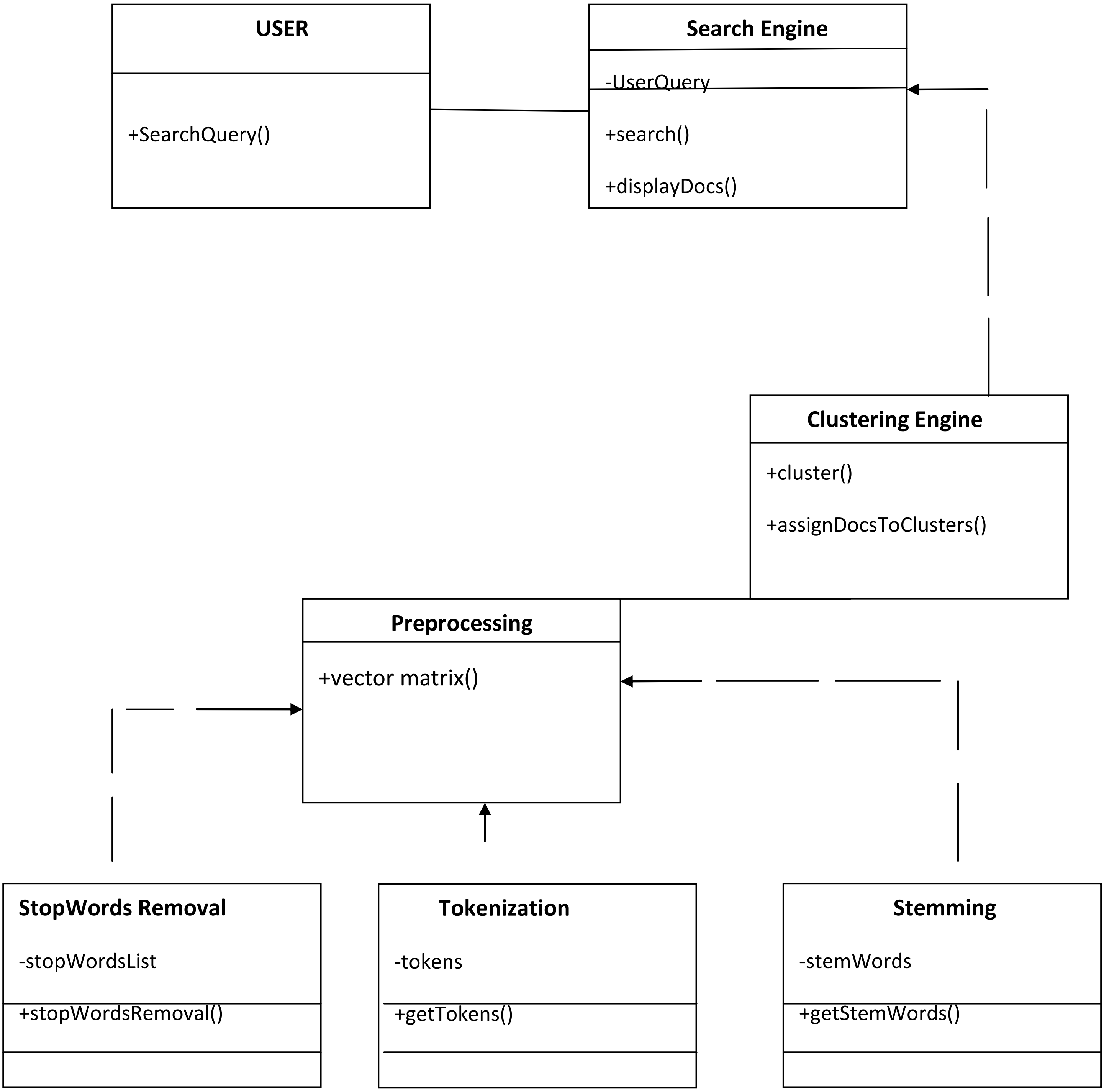


Fig:12 Class Diagram

User: User class has the following operations to be performed.

Operations

searchQuery(): User supplies search query to search engine.

SearchEngine: Search Engine class has the following attributes operations to be performed.

Attributes

userQuery: Of data type string.

Operations

search(): It searches documents to the search query given by the user.

displayDocs(): It displays the documents to the clustering engine.

Clustering Engine: Clustering Engine class has the following operations.

Operations

Cluster:Clustering engine constructs cluster .

assignDocsToClusters(): Clustering engine assign all related documents to the clusters based on similarity measure.

displayResult(): Clustering engine finally displays cluster results to the user.

Preprocessing: This class has the following attributes and operations.

Operations

VectorRepresentation(): forms vector matrix.

Stopwords Removal: This class has the following attributes and operations.

Attributes

List: Of data type string array.

Operations

removeStopWords(): remove the words that are not having much importance by applying

stop list algorithms.

Tokenization: This class has the following attributes and operations.

Attributes

tokens: Of data type string array.

Operations

getTokens(): This method gives all the words in a document by taking some delimiters as separators.

Stemming: This class has the following attributes and operations.

Attributes

stemWords: Of data type string array.

Operations

getStemWords(): This method uses the porter stemmer algorithm to form stemmed words. By following this design the system is implemented.

4.2.2 Activity Diagram

It is a special kind of state chart diagram that shows the flow from activity to activity within a system. It shows the dynamic behavior of the system.

Action States involved in the above activity diagram,

- User-Query
- Giving Results to Clustering Engine
- Preprocessing
- Clustering
- Assign Documents to Clusters
- Display output to User

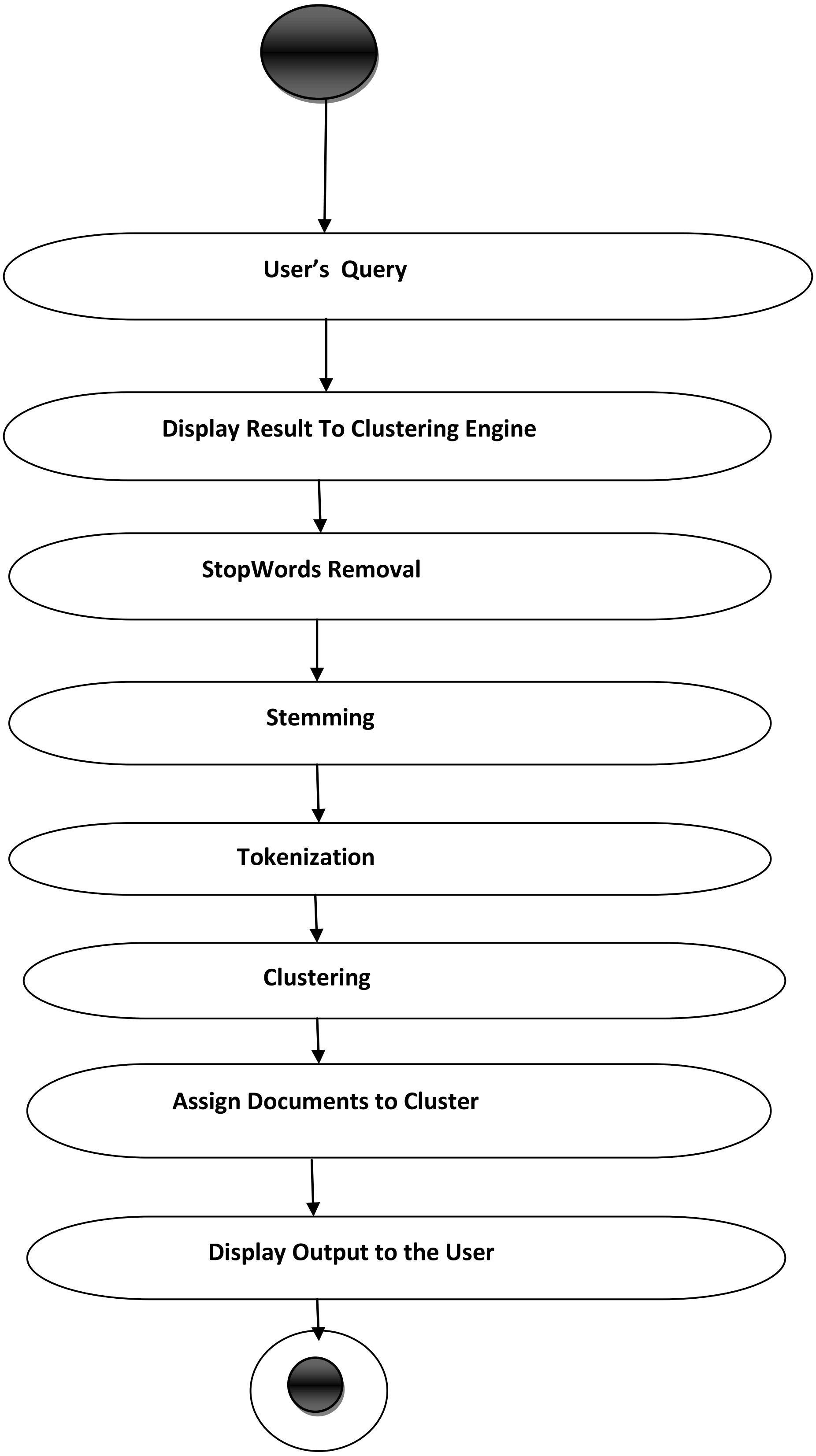

User’s Query
Display Result To Clustering Engine
StopWords Removal
Stemming
Tokenization
Clustering
Assign Documents to Cluster
Display Output to the User

Fig:13 Activity Diagram

Flow of control

1. User supplies the query to the conventional search engine.

2. Search engine displays the results to the clustering engine.

3. Then clustering engine applies preprocessing techniques to the results.

4. Preprocessing involves stop-words removal, tokenization, stemming and finally forms reduced vector matrix representation.

5. Clustering engine constructs cluster hierarchy from reduced binary vector matrix.

6. Clustering engine assign related documents to clusters.

7. Finally clustering engine displays the output to the user.

4.2.3 Sequence Diagram

It emphasis the time-ordering of messages. Every sequence diagram captures the behavior of a single use case.

Objects involved in the above sequence diagram,

- User
- Cluster Engine
- Search Engine

Flow of control

1.User supplies the query to the conventional search engine.

2. Search engine displays the results to the clustering engine.

3. Then clustering engine applies preprocessing techniques to the results.

4.Preprocessing involves stop-words removal, tokenization, stemming and finally forms reduced vector matrix representation

5. Clustering engine constructs cluster hierarchy fromreduced binary vector

matrix.

6. Clustering engine assign related documents to clusters.

7. Finally clustering engine displaysthe output to the user.

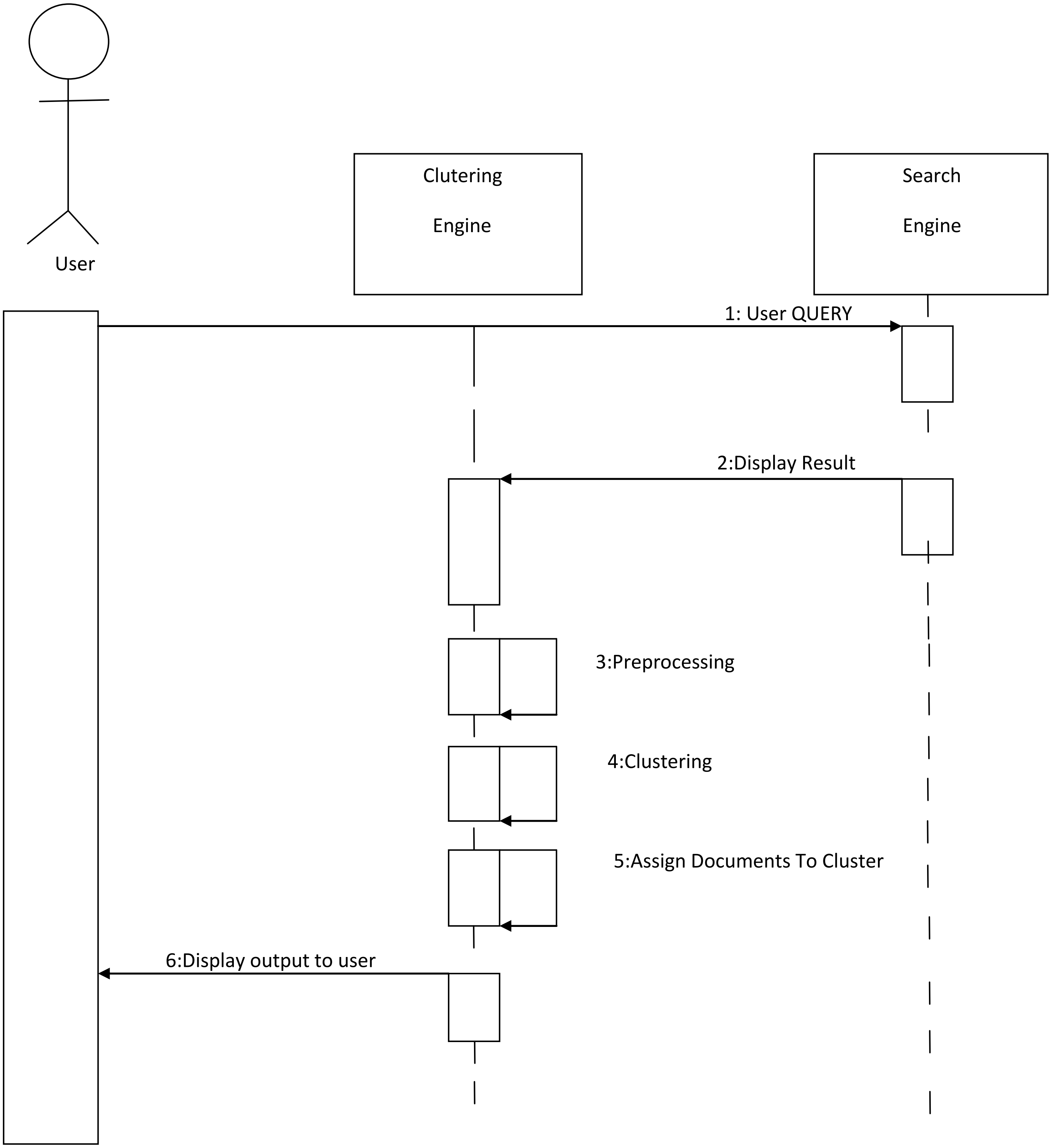


Fig:14 Sequence Diagram

4.2.4 State Transition Diagram

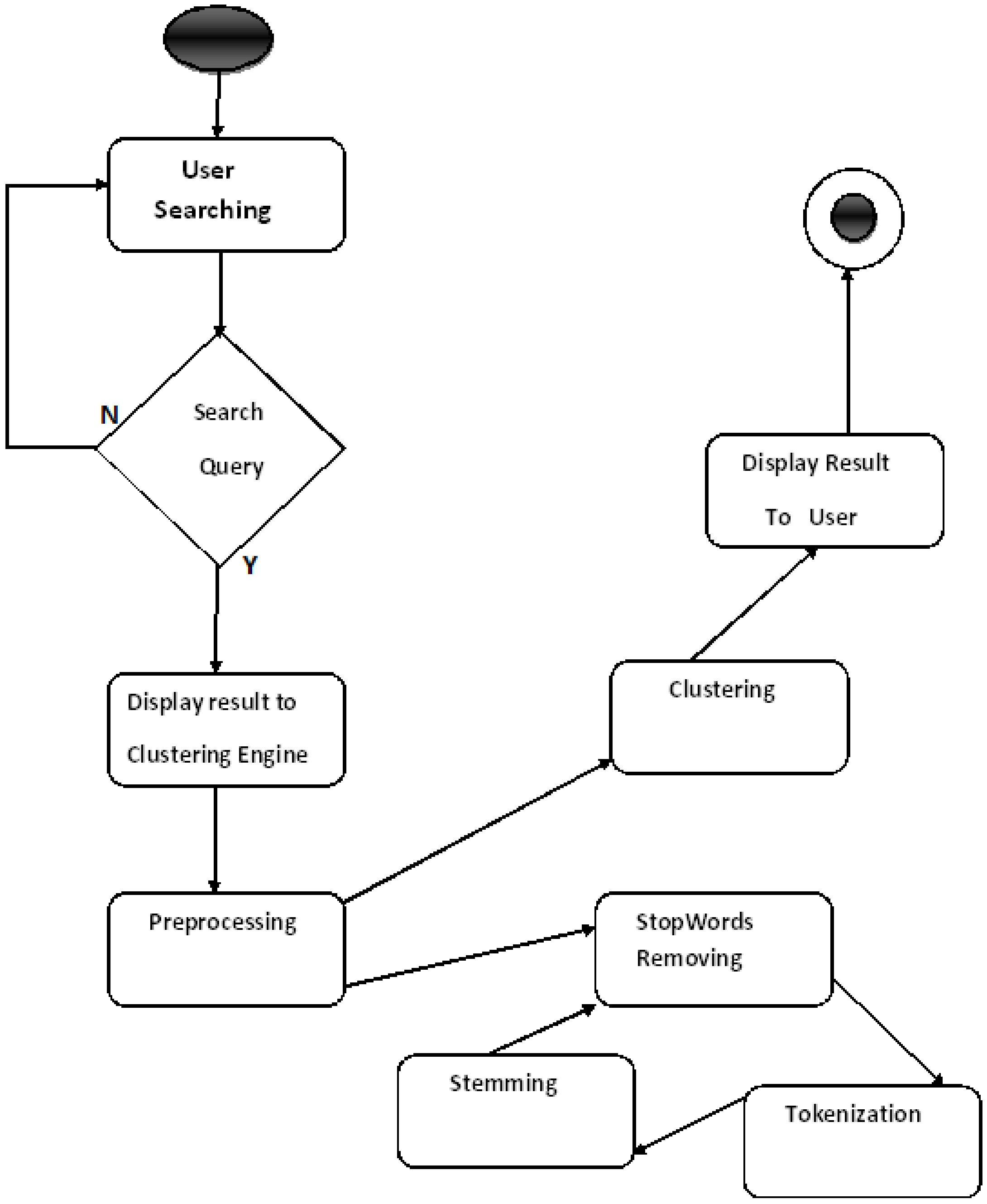


Fig:15 State Transition Diagram

4.2.5 Use Case Diagram

It shows the set of use cases and actors and their relationships. It shows the static view of the system and used in organizing and modeling the behaviors of the system.

Actors involved in above use-case diagram

- User
- Tool

1.User supplies a search query to conventional search engine.

2. Tool takes search engine result as input process that result by applying preprocessing techniques like stemming, stop-words removal, tokenization and forms the binary

3.vector representation of all documents. It finally constructs the cluster , assign the documents to related clusters and display output to the user.

Functionality of Use-Cases

- Search Engine: It provides ranked list of Documents related to the user search query.
- Preprocessing: It includes other use cases like stop-words removal, stemming,

tokenization.

- Binary vector matrix representation: It represents all the documents returned by

search query in binary vector form.

- Reduced binary vector matrix representation:It removes the duplicate rows in binary

vector matrix and gives the reduced binary vector matrix representation.

- Construct Cluster Hierarchy: It gives the representation of clusters in a hierarchical

way by using inheritance concept.

- Assigning Documents to Clusters: It assigns related documents to the clusters

according to similarity measure.

- Display the Output the User: It displays output un the form of hierarchically arranged clusters by displaying words associated with each cluster.

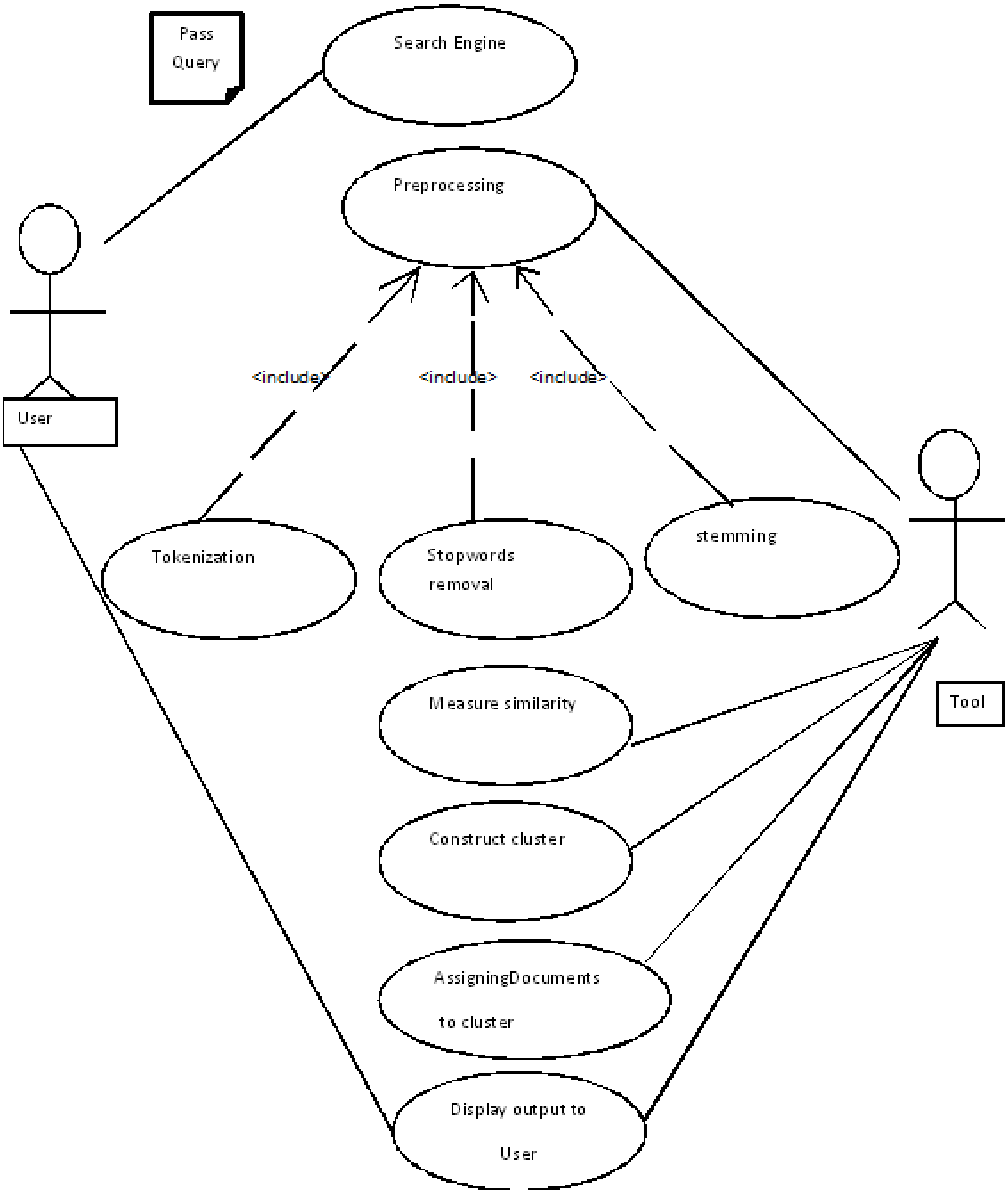


Fig:16 Use Case Diagram

4.2.6 Deployment Diagram

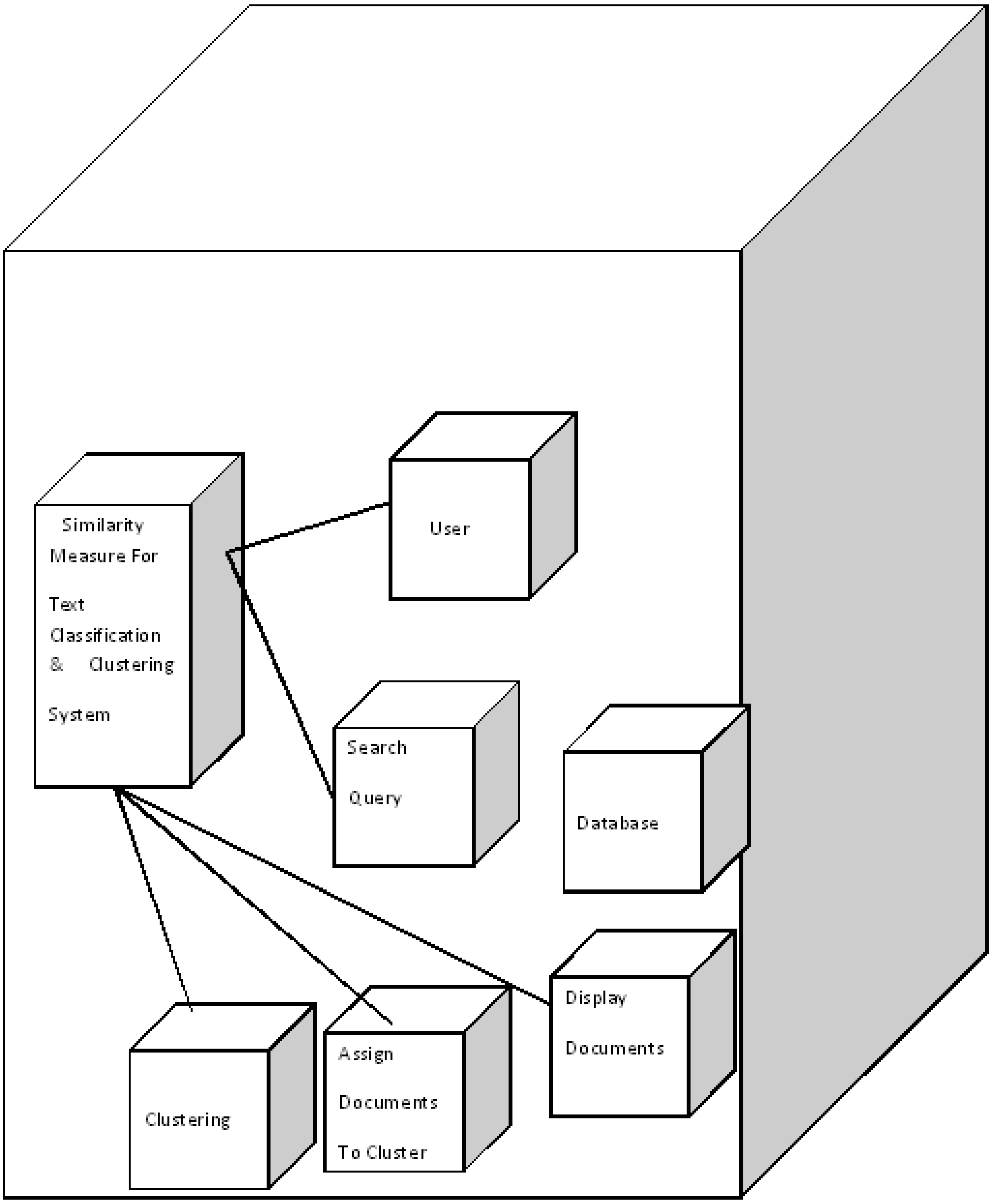

Similarity Measure For Text Classification & Clustering System
User
Search Query
Database
Display Documents
Clustering
Assign Documents To Cluster

Fig:17 Deployment Diagram

5 TECHNICAL SPECIFICATIONS

5.1 ADVANTAGES

- Helping users to effectively navigate, summarize and organize the information.
- Efficient Information retrieval, data mining, and web search.
- Improve the clustering accuracy and less classification time.
- Extracting propositional knowledge-base information from the Web.

5.2 APPLICATIONS

- Document retrieval and text mining.
- Useful in News Websites .
- In Historical saving.
- Getting the Related News .
- Scientific data exploration (e.g. bioinformatics)

6 Bibliography:

[1] Yung-Shen Lin, Jung-Yi Jiang and Shie-Jue Lee,” A Similarity Measure for Text Classification and Clustering”, IEEE Transactions On Knowledge And Data Engineering, 2014.

[2] Gaddam Saidi Reddy and Dr.R.V.Krishnaiah,” Clustering Algorithm with a Novel Similarity Measure”, IOSR Journal of Computer Engineering (IOSRJCE), Vol. 4, No. 6, pp. 37-42, Sep-Oct. 2012.

[3] Shady Shehata, Fakhri Karray, and Mohamed S. Kamel, “An Efficient Concept-Based Mining Model for Enhancing Text Clustering”, IEEE Transactions On Knowledge And Data Engineering , Vol. 22, No. 10, October 2010.

[4] Anna Huang, Department of Computer Science, The University of Waikato, Hamilton, New Zealand,” Similarity Measures for Text Document Clustering”, New Zealand Computer Science Research Student Conference (NZCSRSC), Christchurch, New Zealand, April 2008.

[5] H. Chim and X. Deng, “Efficient phrase-based document similarity for clustering”, IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 9, pp. 1217 – 1229, 2008.

[6] Yanhong Zhai and Bing Liu, “Web Data Extraction Based on Partial Tree Alignment”, International World Wide Web Conference Committee (IW3C2), ACM 1-59593-046, 9/05/2005.

[7] I. S. Dhillon, J. Kogan and C. Nicholas, “ Feature Selection and Document Clustering”, In Berry MW Ed. A Comprehensive Survey of Text Mining, 2003.

[8] Syed Masum Emran and Nong Ye, “Robustness of Canberra Metric in Computer Intrusion Detection”, IEEE Workshop on Information Assurance and Security United States Military Academy, West Point, NY, 5-6 June, 2001.

[9] Alexander Strehl, Joydeep Ghosh, and Raymond Mooney,” Impact of Similarity Measures on Web-page Clustering”, Workshop of Artificial Intelligence for Web Search, July 2000.

[10] S. Kullback and R. A. Leibler, “On information and sufficiency”, Annuals of Mathematical Statistics, Vol. 22, No. 1, pp. 79–86, March 1951.

[11] Mei-Ling Shyu, Shu-Ching Chen, Min Chen and Stuart H. Rubin,” Affinity-Based Similarity Measure for Web Document Clustering”, Distributed

Multimedia Information System Laboratory, School of Computer Science Florida International University Miami, FL 33199, USA.

[12] J. Han and M. Kamber,Data Mining: Concepts and Techniques,2nded. San Francisco, CA, USA: Morgan Kaufmann; Boston, MA,USA: Elsevier, 2006.

ANNEXURE

ANNEXURE A

Abbreviations and Acronyms

DFD: Data Flow Diagram

UML: Unified Modeling Language

RAM: Random Access Memory

STD: Suffix Tree Document

VSD: Space Document

HAC :Hierarchical Agglomerative Clustering algorithm

Annexure B

Laboratory Assignment No. 1

Aim:

Refer Chapter 7 of first reference to develop the problem under consideration and justify feasibility using concepts of knowledge canvas and IDEA Matrix

Canvas:

Project Canvas is a visual tool that improves communication in project teams and provides a simplified project overview. Project Canvas is the missing link when it comes to creating a red thread in our projects. Project Canvas is a great guideline, as it quickly maps out the vital elements of a project.

Participants	Goals	User Benefits	Activities	Deliverables
1.Rahul Nalawade 2.Akash Samal	Efficiently measure the similarity Between the articles and perform clustering on it One article is display at the top and other Similar articles at the bottom So duplication is avoided.	This project can help user for getting the Related news. User do not need to go for every news site for ananalyse &read it.all the news are present in cluster	List the concrete tasks and actions the team will take to the project goals PHASE 1: Concept and Design-Create a design concept for the relevant pages PHASE 2: Testing -Evaluate the new design in a qualitative user test -Plan and conduct tests	Indicate the outcomes and documents that will be shown to users. This does not include working documents, project plan and similar Concept & Design -Final Updated Designs Testing -Final Report Implementation will be done

Risks	Milestones	Constrains	Scope
Identify possible future events that could have a negative impact on project. RISK: If video format does not suitable to summarized, system fails	July'15 –group formation & selection of domain Aug-sept'15 – training,anlysis oct – designing Dec- coding March- testing,documentation	Only text classification and Clustering is done	The scope of the project is limited to Bollywood news but it can be extended to other news also. The project takes number of articles from different sources

Table : Project Canvas

IDEA Matrix:

A Similarity Measure For Text classification &Clusterring	Who did this before?	How did they do it?	What makes mine better?
Technical	Yung-Shen Lin, Jung-Yi Jiang, and Shie-Jue Lee presented a novel similarity measure between documents. To improve the efficiency, we have provided an approximation to reduce the complexity involved in the computation	algorithms used k-NNsingle-label classification, k-NN based multi-label classification,k-means clustering,and hierarchical agglomerative clustering (HAC).	A Similarity Measure For Text classification &Clusterring is a useful technique that can be applied for clustering text data. We use improved k-means clustering algorithm
Aesthetic	usefulness of a similarity measure could depend on application domain, featureformats, classification/clustering algorithms	For aesthetic of old technology they just developed normal GUI with lots of menus and functions which makes it difficult for user to understand and they had implemented some well and good aesthetic which looks bit beautiful.	It has very simple GUI. This makes it easy to use for non technical user. It takes itself beyond document clustering.

Narrative	first case, the similarity increases as the difference between the two involved feature values decreases. Furthermore, the contribution of the difference is normally scaled. For the second case, a fixed value is contributed to the similarity. For the last case, the feature has no contribution to the similarity	The effectiveness of measure is evaluated on several real-world data sets for text classification and clustering problems	As number of data is generated from the news articles.it is difficult to read data manualy from all news sites.so we developing Similarity measure clustering system. We could use initial boundry constraint on k value or use optimization algorithm to minimize cross validation error

Table 9.2: IDEA Matrix

Here is the matrix for our idea. Because of the fact that we feel this project would do best in a video summarization techniques, we believe it would create meaningful video summary.

Laboratory Assignment No. 2

Aim: Project problem statement feasibility assessment using NP-Hard, NP-Complete or satisfy ability issues using modern algebra and/or relevant mathematical models

P type problem: The Collection of problems that can be solved in polynomial time is called P. This polynomial is small degree. The problems belonging to this class are easy to solve and can be solved using tractable input.

NP type problem: The algorithm in which every operation may not have unique result, rather there can be specified set of possibilities for every operation. Such an algorithm is called non-deterministic algorithm. Non-deterministic algorithm is a two stage algorithm.

A. Non-deterministic (guessing) stage: Generate an arbitrary string that can be thought of as a candidate solution.

B. Deterministic (Verification) stage: In this stage, it takes as input the candidate solution and the instance to the problem and returns yes if the candidate solution represents actual solution.

There are two types of NP-type problem:

A. **NP-complete**: They are harder to compute rather than to verify; they could not be solved in polynomial time but they can be verified in polynomial time.
B. **NP-hard:** They need not have solutions verifiable in polynomial time.
C. For the present methodology The input given to these methods is number of articles. So as input is infinite, output will be infinite and hence methods are Non deterministic.The time required for computation will depend on total no of words in articles. Therefore, this problem statement comes under NP-Complete.

Mathematical Model:

Let S be the solution perspective of the class

S={s, e, i, o, f, DD, NDD, success, failure}

s=Start of program

e = the end of program

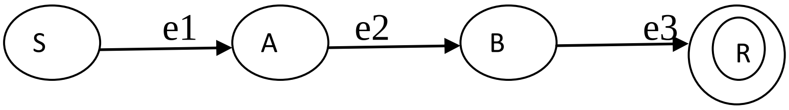
i= the news or article from different sources

o=percentage of articles similarity,cluster of similar articles

Success-Clustring is done

Failure- error is displayed to the user

Computational Model



Where,

S={Start state}

A={Preprocessing()}

B={Clutering()}

R={Final Result}

Laboratory Assignment No. 3

Aim-

Use of divide and conquer strategies to exploit distributed/parallel/concurrent processing of the above to identify objects, morphisms, overloading in functions (if any), and functional relations and any other dependencies (as per requirements).

Concept-

A divide and conquer algorithm works by recursively breaking down a problem into two or more sub-problems of the same (or related) type (divide), until these become simple enough to be solved directly (conquer). The solutions to the sub-problems are then combined to give a solution to the original problem.

System Specifications and dependencies-

Our proposed system allows users to use the same credentials for logging into different applications/accounts. The system makes use of Kerberos for implementing the single sign on system and the application needs to be Kerberos enabled so that the user is not asked for username and password every time he needs to log into a particular account.

The system comprises of following major components-

- Document retrieving(web page crawling)
- Preprocessing

  Stop words Removal

  Tokenization

  Stemming
- Load dataset
- Clustering

- So this project is used by everyone with the help of GUI which we provide. It is very easy to use.

In computer science, **Divide and Conquer (D&C)** is an algorithm design paradigm based on multi-branched recursion. A divide and conquer algorithm works by recursively breaking down a problem into two or more sub-problems of the same (or related) type (**divide**), until these become simple enough to be solved directly (**conquer**). The solutions to the sub-problems are then combined to give a solution to the original problem. This divide and conquer technique is the basis of efficient algorithms for all kinds of problems, such as sorting (e.g., quick sort, mergesort), multiplying large numbers syntactic analysis (e.g., top-down parsers), and computing the discrete Fourier transform (FFTs).

Advantages of Divide and Conquer:-

- Solving difficult problems.
- The divide-and-conquer paradigm often helps in the discovery of efficient algorithms.
- Parallelism - Divide and conquer algorithms are naturally adapted for execution in multi-processor machines.
- Divide-and-conquer algorithms naturally tend to make efficient use of memory caches.

Laboratory Assignment No. 4

Aim:-

Use of above to draw functional dependency graphs and relevant Software modelling methods, techniques including UML diagrams or other necessities using appropriate tools.

Concept-

Functional Dependency Graph:

A functional dependency is a relationship between two attributes. Typically between the PK and other non-key attributes within the table. For any relation R, attribute Y is functionally dependent on attribute X (usually the PK), if for every valid instance of X, that value of X uniquely determines the value of Y.

X ———> Y

A functional dependency is a connection between two sets of attributes in a relation from a database. Functional dependency diagrams are useful diagrammatic way of showing functional dependencies.

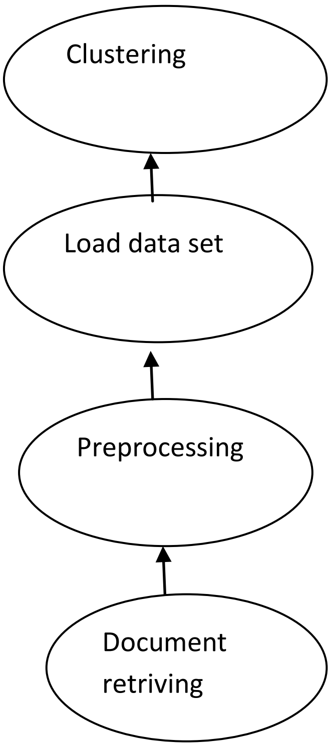


Figure : Functional Dependency Graph

UML Diagrams

The unified modeling language is standard language for specifying,Visualizing,Constructing and documenting the software system and its components. It is a graphical language which provides a vocabulary and set ofsemantics and rules. The UML focuses on the conceptual and physical representation of the system. It captures the decisions and understandings about systems that must be constructed. It is used to understand, design, configure, maintain and control Information about the systems

The vocabulary of the UML encompasses 3 kinds of building blocks.

They are

- Things
- Relationships
- Diagrams

Diagrams in the UML :

Class Diagram

Activity Diagram

Sequence Diagram

Use-Case Diagram

Deployment Diagram

In our project we designed our system by using the following Diagrams.

Class Diagram

It shows a set of classes, interfaces and collaborations and their relationships. It addresses the static process, view of the system Classes involved in the above class diagram,

- User
- Search Engine
- Clustering Engine
- Preprocessing
- Stop-words Removal

- Tokenization
- Stemming

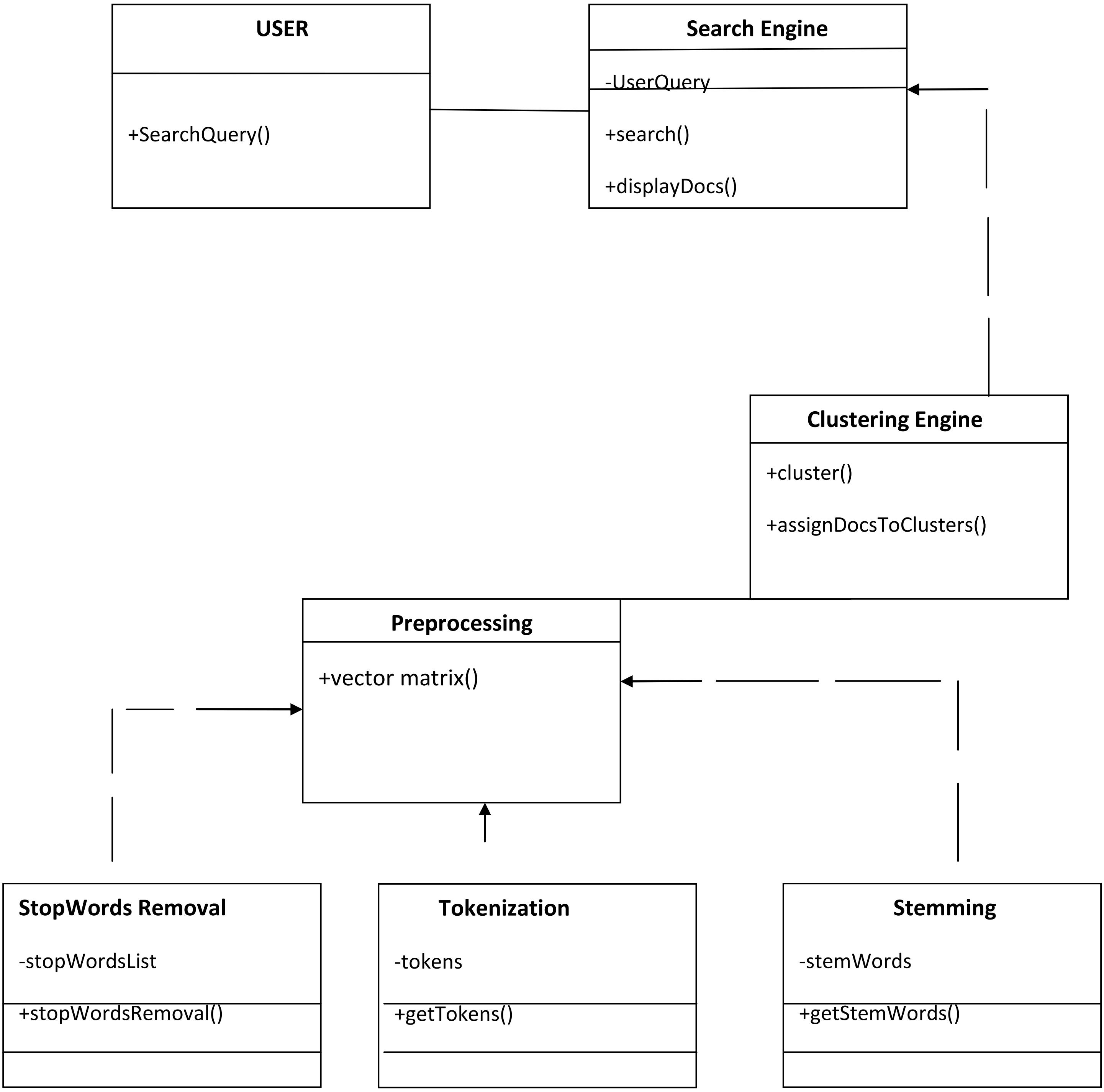


Fig:Class Diagram

User: User class has the following operations to be performed.

Operations

searchQuery(): User supplies search query to search engine.

SearchEngine: Search Engine class has the following attributes operations to be performed.

Attributes

userQuery: Of data type string.

Operations

search(): It searches documents to the search query given by the user.

displayDocs(): It displays the documents to the clustering engine.

Clustering Engine: Clustering Engine class has the following operations.

Operations

Cluster:Clustering engine constructs cluster .

assignDocsToClusters(): Clustering engine assign all related documents to the clusters based on similarity measure.

displayResult(): Clustering engine finally displays cluster results to the user.

Preprocessing: This class has the following attributes and operations.

Operations

VectorRepresentation(): forms vector matrix.

Stopwords Removal: This class has the following attributes and operations.

Attributes

List: Of data type string array.

Operations

removeStopWords(): remove the words that are not having much importance by applying

stop list algorithms.

Tokenization: This class has the following attributes and operations.

Attributes

tokens: Of data type string array.

Operations

getTokens(): This method gives all the words in a document by taking some delimiters as separators.

Stemming: This class has the following attributes and operations.

Attributes

stemWords: Of data type string array.

Operations

getStemWords(): This method uses the porter stemmer algorithm to form stemmed words. By following this design the system is implemented.

Activity Diagram

It is a special kind of state chart diagram that shows the flow from activity to activity within a system. It shows the dynamic behavior of the system.

Action States involved in the above activity diagram,

- User-Query
- Giving Results to Clustering Engine
- Preprocessing
- Clustering
- Assign Documents to Clusters
- Display output to User

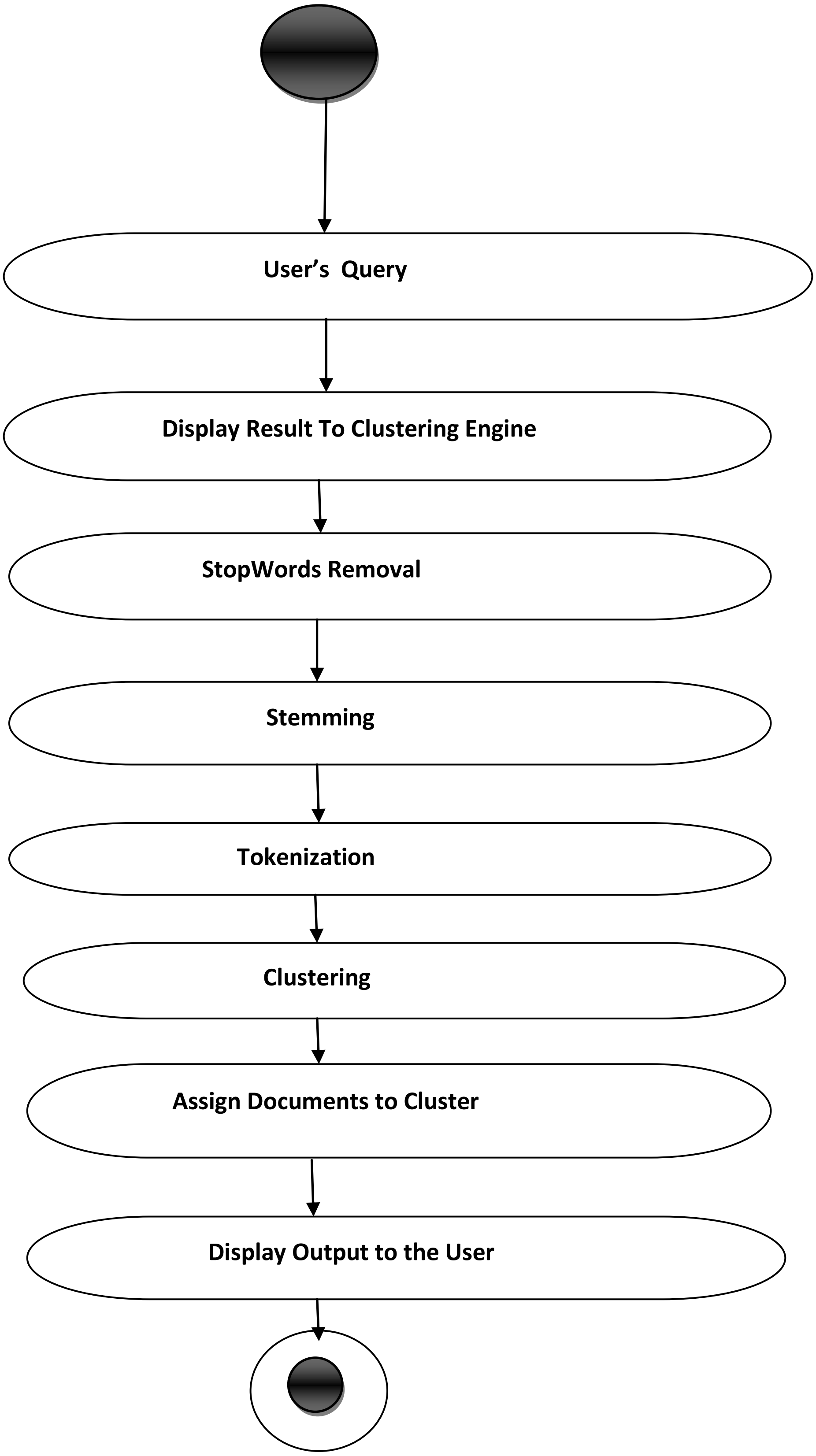

User's Query
Display Result To Clustering Engine
StopWords Removal
Stemming
Tokenization
Clustering
Assign Documents to Cluster
Display Output to the User

Fig: Activity Diagram

Flow of control

1. User supplies the query to the conventional search engine.

2. Search engine displays the results to the clustering engine.

3. Then clustering engine applies preprocessing techniques to the results.

4. Preprocessing involves stop-words removal, tokenization, stemming and finally forms reduced vector matrix representation.

5. Clustering engine constructs cluster hierarchy from reduced binary vector matrix.

6. Clustering engine assign related documents to clusters.

7. Finally clustering engine displays the output to the user.

Sequence Diagram

It emphasis the time-ordering of messages. Every sequence diagram captures the behavior of a single use case.

Objects involved in the above sequence diagram,

- User
- Cluster Engine
- Search Engine

Flow of control

1.User supplies the query to the conventional search engine.

2. Search engine displays the results to the clustering engine.

3. Then clustering engine applies preprocessing techniques to the results.

4.Preprocessing involves stop-words removal, tokenization, stemming and finally forms reduced vector matrix representation

5. Clustering engine constructs cluster hierarchy fromreduced binary vector

matrix.

6. Clustering engine assign related documents to clusters.

7. Finally clustering engine displaysthe output to the user.

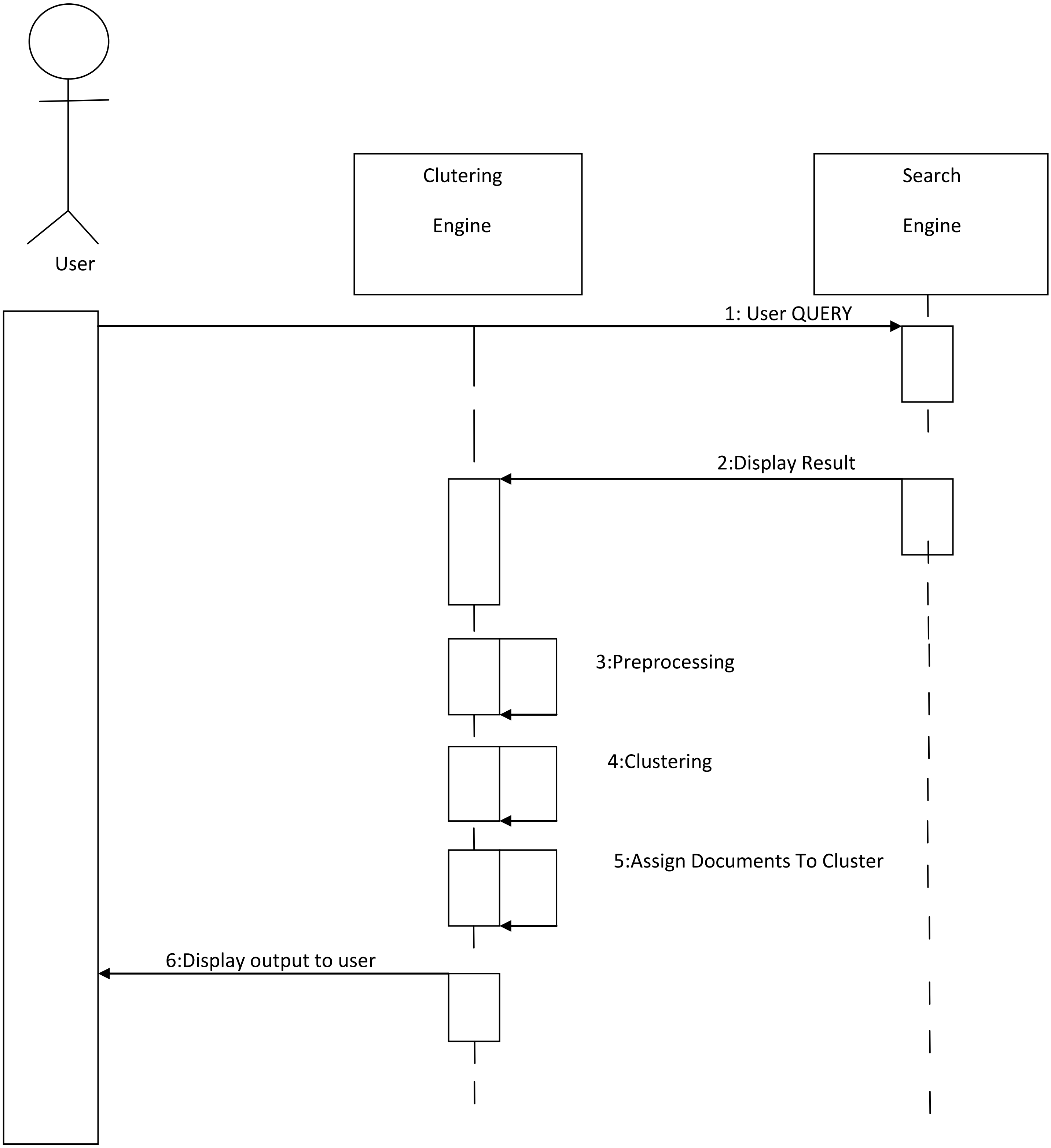


Fig: Sequence Diagram

State Transition Diagram

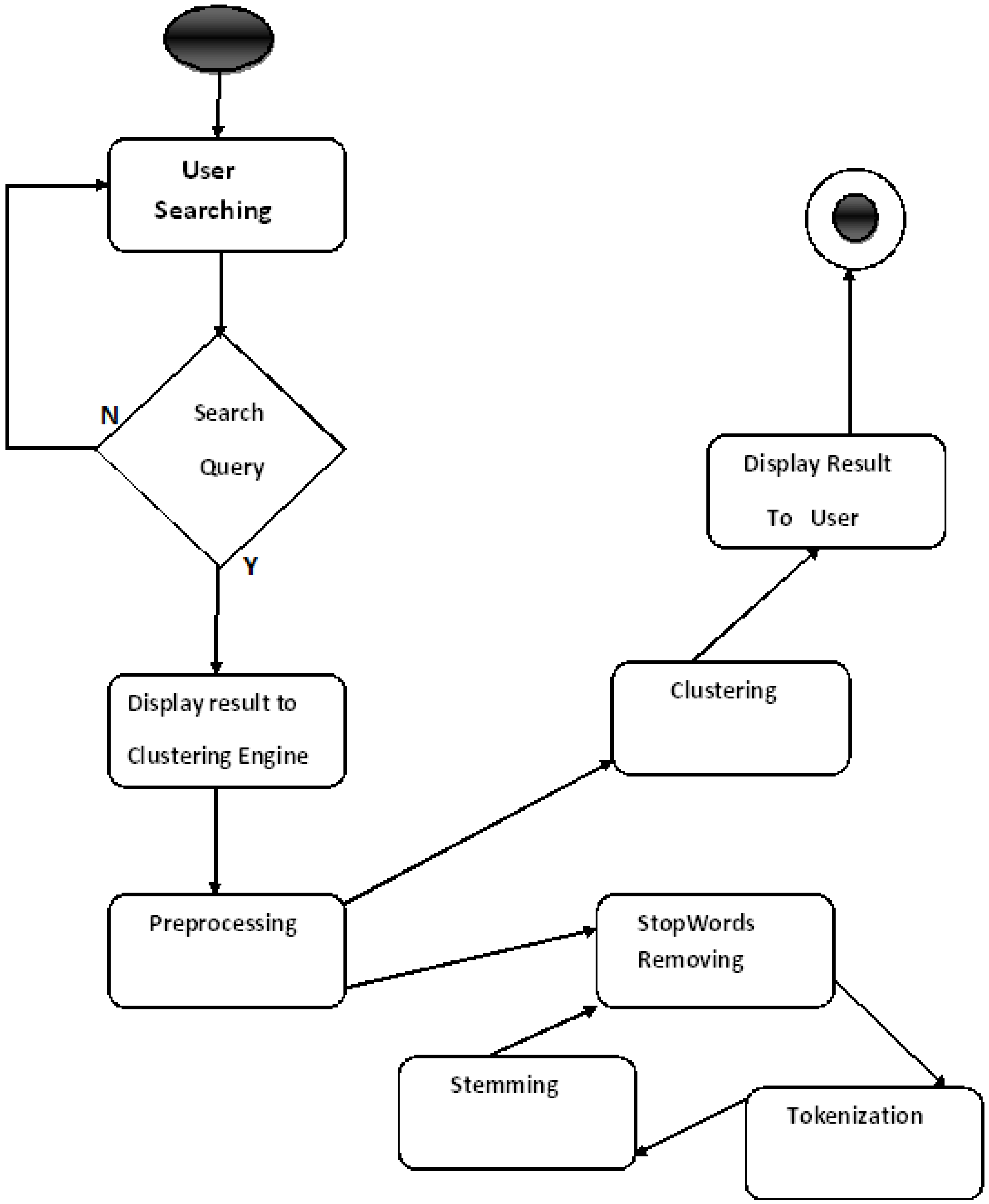


Fig:State Transition Diagram

Use Case Diagram

It shows the set of use cases and actors and their relationships. It shows the static view of the system and used in organizing and modeling the behaviors of the system.

Actors involved in above use-case diagram

- User
- Tool

1.User supplies a search query to conventional search engine.

2. Tool takes search engine result as input process that result by applying preprocessing techniques like stemming, stop-words removal, tokenization and forms the binary

3.vector representation of all documents. It finally constructs the cluster , assign the documents to related clusters and display output to the user.

Functionality of Use-Cases

- Search Engine: It provides ranked list of Documents related to the user search query.
- Preprocessing: It includes other use cases like stop-words removal, stemming,

tokenization.

- Binary vector matrix representation: It represents all the documents returned by

search query in binary vector form.

- Reduced binary vector matrix representation:It removes the duplicate rows in binary

vector matrix and gives the reduced binary vector matrix representation.

- Construct Cluster Hierarchy: It gives the representation of clusters in a hierarchical

way by using inheritance concept.

- Assigning Documents to Clusters: It assigns related documents to the clusters

according to similarity measure.

- Display the Output the User: It displays output un the form of hierarchically arranged clusters by displaying words associated with each cluster.

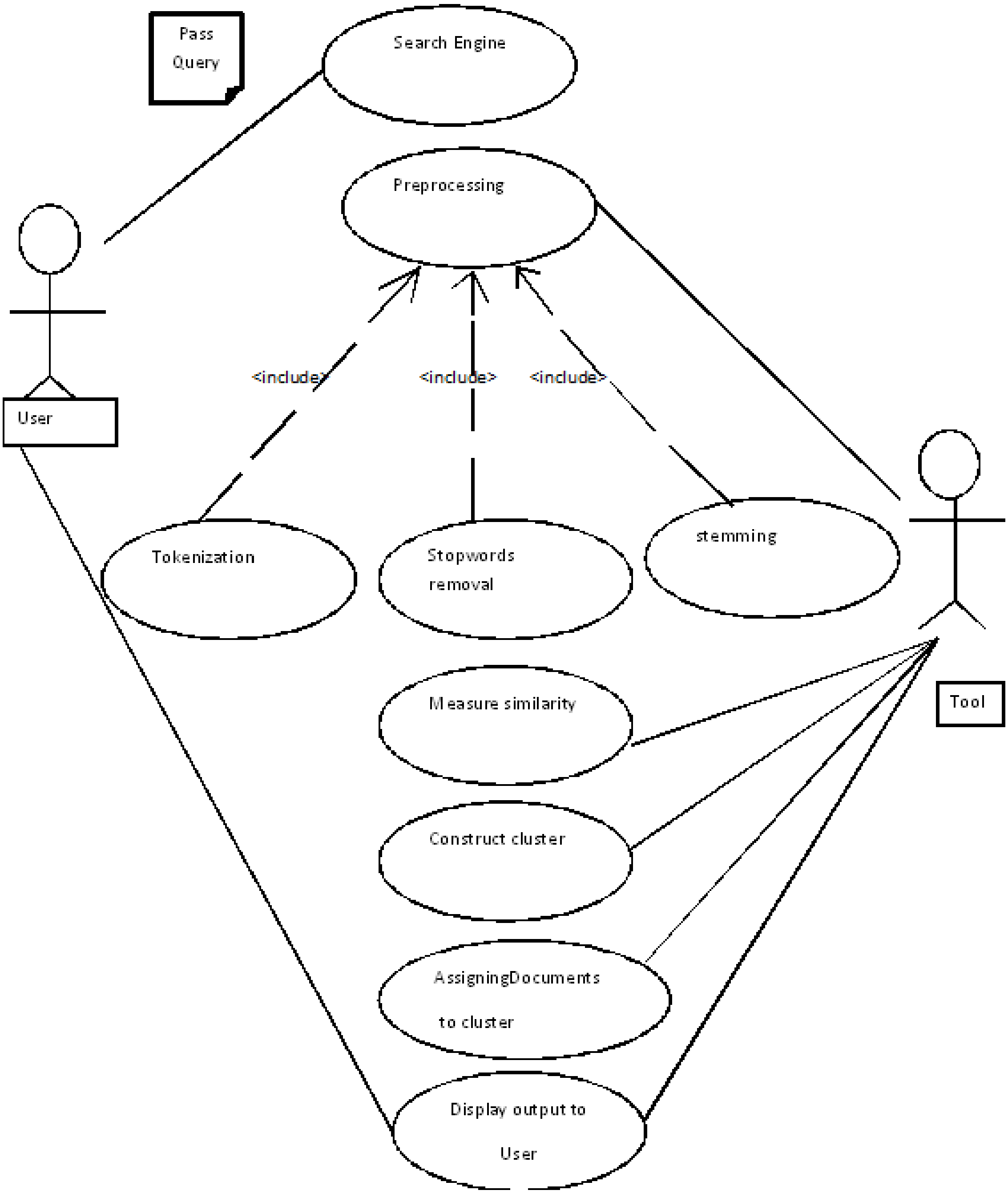


Fig: Use Case Diagram

Deployment Diagram

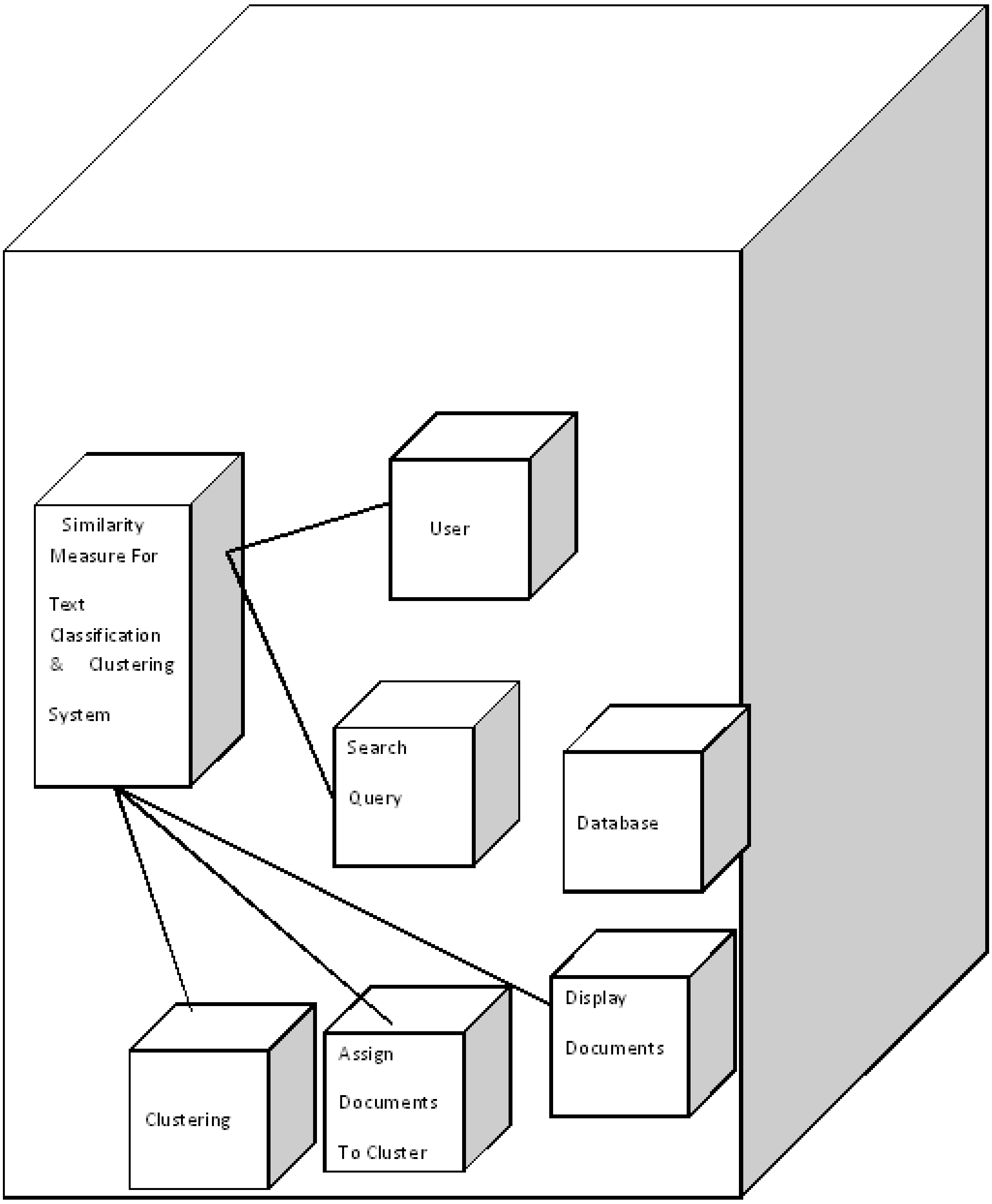

Similarity Measure For Text Classification & Clustering System
User
Search Query
Database
Display Documents
Clustering
Assign Documents To Cluster

Fig:Deployment Diagram

Laboratory Assignment No. 5

Aim-

Testing of project problem statement using generated test data (using mathematical models, GUI, Function testing principles, if any) selection and appropriate use of testing tools, testing of UML diagram's reliability.

Concept-

Now days we can get lots of **Software Testing Tools** in the MARKET. Selection of tools is totally based on the project requirements & commercial (Proprietary/Commercial tools) or free tools (Open Source Tools) you are interested. Off Course, free Testing Tools may have some limitation in the features list of the product, so it's totally based on what are you looking for & is that your requirement fulfil in free version or go for paid Software Testing Tools.

The tools are divided into different categories as follows:

- Test Management tools
- Functional Testing Tools
- Load Testing Tools

Manual Testing-

Manual testing is the process of manually testing software for defeats. It requires a tester to play the role of end user and use most of all features of the application to ensure correct behavior.

The manual testing is very basic type of testing which helps to find the bugs in the application under test. It is preliminary testing, must to carried out to prior to start automating the test cases also needs to check the feasibility automation testing. In manual testing, we are checking whether after clustering, same articles are getting grouped in same cluster

Smoke testing is performed on these modules and if they fail this test, modules are reassigned to respective developers for fix. For passed modules manual testing is carried out from the written test cases. If any bug is found that get assigned to module developer and get logged in bug tracking tool. On bug fix tester do bug verification and regression testing of all related modules. If bug passes the verification it is marked as verified and marked as closed. Otherwise above mentioned bug cycle gets repeated.

Different tests are performed on individual modules and integration testing on module integration. These tests include Compatibility testing i.e. testing application on different hardware, OS versions, software platform, different browsers etc. Load and stress testing is also carried out according to SRS. Finally system testing is performed by creating virtual client environment. On passing all the test cases test report is prepared and decision is taken to release the product!

Testing process include following level:

Level 1 –

Check the user enters correct data to search related articles if not then show error message

Level 2 –

Check the articles are preprocessed correctly.if not then do it again.

Level 3 –

Standard software development and maintenance processes are integrated throughout an organization; a Software Engineering Process Group is in place to oversee software processes, and training programs are used to ensure understanding and compliance.

Level 4 –

Track Project performance Is predictable, and quality is consistently high.

Level 5 –

The focus is on continuous process improvement. The impact of new processes and technologies can be predicted and effectively implemented when required.

Annex C: Plagiarism Report

Annex D: External Feedback

External Comment/Suggestion:

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

…………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

……………………………………………………………………………………………………

Name of the External: ……………………………………………..

Sign: …………………………………………….