"INTEGRATION OF TEXT USING DOCUMENT SIMILARITY WITH ENHANCED K-MEAN CLUSTERING"

A Dissertation Report

Submitted By

ROHINI TEWATIA

(11302434)

Submitted to

Department of CSE/IT

In partial fulfillment of the requirements for the award of the Degree of

Master of Technology in Computer Science Engineering

Under the guidance of

Ms. Sandeep Kaur

Assistant Professor,

School of Computer Science, LPU

(MAY 2015)

ABSTRACT

Integration of data from multiple sources to get useful information is the main challenge for an organization. There are many fields in which integration plays a crucial role. Data integration refers to the process in which data is combined from multiple sources to provide unified view of data to the users. The process of integration is transparent from the users.

In this experimental work, we will propose methodologies for integrating large numbers of Text data sets based on clustering. In our research work we use the enhance version of K-Mean and UPGMA algorithms with Euclidean distance i.e. K-Mean and Hybrid approach with cosine similarity for integrating the texts from diverse sources. Finally we have done comparison of existing and proposed clustering methods and selected the best technique for integrating the text.

CERTIFICATE

This is to certify that **Ms. Rohini Tewatia** has completed M.Tech Dissertation proposal titled **"Integration of Text Using Document Similarity with Enhanced K-Mean Clustering"** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation proposal has submitted for any other degree or diploma. The dissertation proposal is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech Computer Science & Engg.

Date:

Signature of Advisor

Name: Sandeep Kaur

UID:

ACKNOWLEDGEMENT

I would like to present my deepest gratitude to **Ms. Sandeep Kaur,** Assistant Professor (Department of Computer Science) for her guidance, advice understanding and supervision throughout the development of this Dissertation study. I would like to thank to the **Project Approval Committee members** for their valuable comments and discussions. I would also like to thank to **Lovely Professional University** for the support on academic studies and letting me involve in this study.

DECLARATION

I hereby declare that the Dissertation proposal entitled "**Integration of Text Using Document Similarity with Enhanced K-Mean Clustering"** submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: _____

Investigator: Rohini Tewatia

Registration No. 11302434

TABLE OF CONTENT

List of Tables

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

DATA MINING

Data mining is the process through which we can generate the useful information from large amount of data. Data Mining is also known as KDD, which means "Knowledge discovery in database". KDD is the process designed to generate patterns that show the well-defined relationship among data. This is also called knowledge discovery from the large amount of database. The goal of data mining is to find the useful and suitable knowledge from the data. For example, we have lot of data on internet, but not all the data is useful for us and we choose only that data which is important for us. We search only that website on the web, which helps us in finding the data required by us.

The knowledge discovery in data can be achieved by following the steps given below:

1. **Data Cleaning:** In this step, the irrelevant and noisy content available in data has removed in order to get more effective and accurate results.
2. **Data Integration:** In this step, the different types of data from multiple data sources have combined at a common source.
3. **Data Selection:** In this step we will select the data which is relevant for the mining.
4. **Data Transformation:** In this stage, the selected data has changed into accurate format for the procedure of data mining.
5. **Data Mining:** This is the important step in which the techniques is using to extract the pattern from data.
6. **Pattern Evaluation:** In this step patterns have been discovers after the mining of the data.
7. **Knowledge Representation:** This is the final step in which knowledge is visually represented to the users. Knowledge representation use visualization techniques to help understanding of user and taking the output of the KDD**.**

Data mining is playing a vital role in many of the field such as market-basket analysis, classification, etc. In data mining, frequent item sets have significant role which is used to find out the correlations between the fields of database.

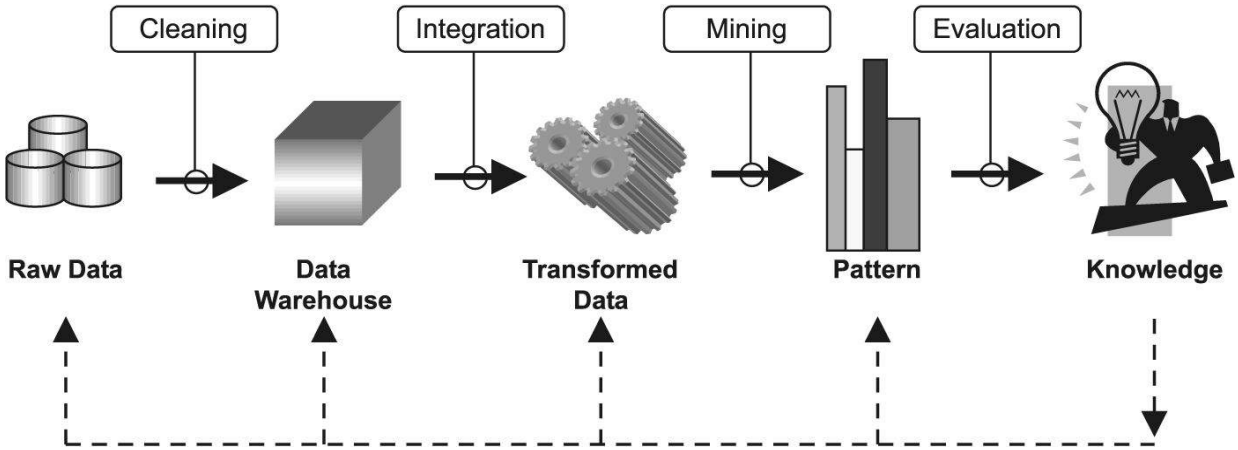

Figure 1.1 Knowledge discovery in Database

1.1 DATA INTEGRATION

Data Integration in Data Mining:

Data Integration is the process, which comes under data processing. Data integration formally defined by < G, S, M > where as G is the global schema, S is the set of sources and M is the mapping between G and S. In the process of Data Integration useful information from multiple sources are combining at one place. It is difficult to collect the set of data from different database at a common place, because data at different sources are available in different formats. After that, it merges the data in unified view for the users. In data integration, we apply mathematical calculation and equation for the computation. Data integration helps in reducing duplication in the combined data. Integration the data from multiple sources help in mining process, which leads to present the more accurate information to the users. The main challenge in the data integration is because of relational database. Since websites have integrated with social sites, such as FB, Twitter, Google+, LinkedIn, etc. because of the social websites, organizations will come to know about their customers their favorite links, their demands and integration will also help the organizations to know about their current trends in the market competition. However, due to the constraint of relational data model, it requires parallel

efforts of modifying database schema design in organize to collect information of FB or twitter's conversations and actions from page viewer. Data integration conceals the obstacle of such problem. It gives the user unique and apparent idea of the information.

There are four types of heterogeneity in the data integration:-

1) **Architectural Heterogeneity:-**This type of heterogeneity occurs due to the incompatibility with hardware resources, operating system and network.

2) **Structural Heterogeneity**: - This type of heterogeneity occurs when we deal with different data models such as relational database or object oriented database.

3) **Syntactical Heterogeneity**: - This type of heterogeneity occurs when we deal with different language and data representation.

4) **Semantic heterogeneity**: - This type of heterogeneity occurs when meaning of the information is not representing properly.

Among all these heterogeneity, semantic heterogeneity is difficult to solve. Semantic heterogeneity occurs when different parties create datasets for the same domain independently.

To avoid redundancy and inconsistency among datasets data integration should be done carefully.

There are Numbers of Issues in Data Integration

1) **Entity identification problem: -** When different parties create datasets for the same domain independently which can lead to entity identification problem. For example customer attribute created for the same domain but independently, so how the data analyst will come to know this customer attribute is same but with the different identification [1].
2) **Redundancy: -** Redundant is also main issue in the data integration. Redundant can be termed as when same attribute is derived from another attribute. Redundant can be determined by correlation analysis. Correlation analysis based on the available data measure how the one attribute strongly dependent to another attribute. It is denoted by χ^2 [1].

EQUATION: -

$$\chi^2 = \sum \frac{(O-E)^2}{E} \quad [1]$$

Whereas O is the observed frequency and E is the expected frequency [1].

In the previous years, ETL was ideal in data integration strategy because database did not have the power to deal with the complex data. In the process of the ETL, data is first extract from the different databases and after that, data is moved to intermediate platform i.e. transform in which data is converted into accurate format and then data is load from source to destination.

.

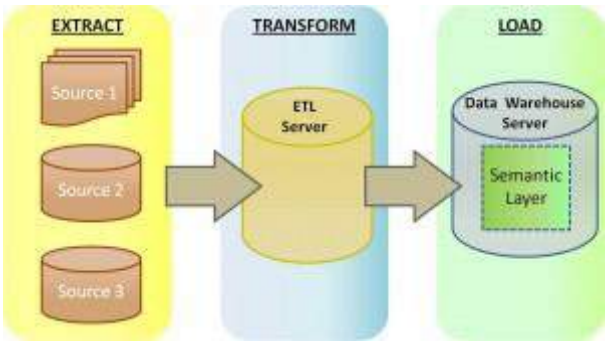

Fig 1.2 ETL (Extract, Transform and Load)

Extract: - In this step, data will be extract from the different sources and connect to the source systems in which necessary data will be select and collect for the processing within the data ware house. Complexity of extract phase is directly proportional to the amount of data and the type of data is extracting.

Transform: - In this step on the extracted data, rules and functions will be applied to convert it into accurate form. If the good quality of data has been extracted then need less transformation on that data. If the data is inconsistent then filtering, sorting can be applied to make the data in consistency form.

Loading: -This is the last step in ETL process, in this data is loaded into the target database or data ware house.

Disadvantages of ETL

- To implement ETL tool additional hardware required.
- Flexibility of ETL decreases because it dependency on the ETL tool vendor.

1.1.1 CLUSTERING

Clustering technique is used for the unsupervised learning in which data elements do not have the label information on it. In cluster, unlabeled datasets will produce true model in the form of result. Cluster is form of learning by observation where as classification is form of learning by experience. Cluster is the process of partitioning the set of objects into cluster such that similar objects will cluster into one group and those objects which are not similar will be in another group [1].

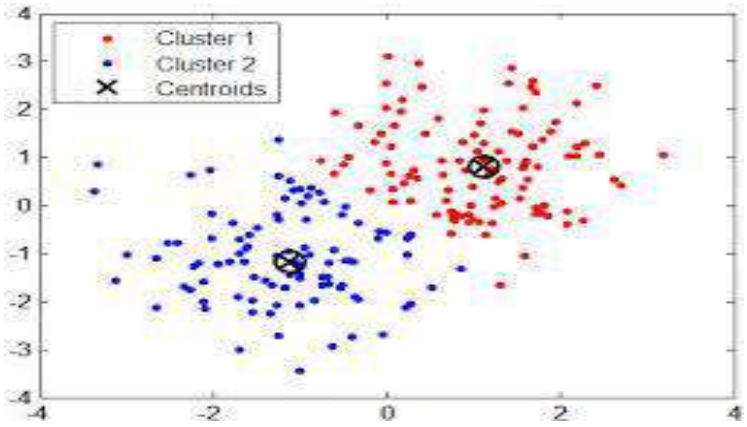

Fig1.3 Representation of clusters

1.1.1.2 CLUSTERING METHODS

There are three clustering methods:

1) **Partitioning Method**: This is the simple and most popular method of cluster analysis in which set of objects is partitioned into clusters. Partitioning method is the distance based method. This method provides good result when the data sets

are small or medium in size. To represent the cluster center, it uses the Mean and Medoid.

2) **Hierarchical method**: Hierarchical method is also known as connectivity based clustering. In this method objects are connected based on their distance to form the cluster. According to the distances, different clusters can be formed by using dendrogram.
3) **Grid-Based method**: Multi resolution grid structure approach is used in the grid based method. Fast processing time is the advantage of this method. Grid-based method independent from the number of the data objects but depend upon grid size. Examples STING (statistical information grid) and CLIQUE (clustering in Quest) [1].

Methods	**Characteristics**
Partitioning Method	- To represent the cluster center, it uses the Mean and Medoid. - Provides good result when the data sets are small or medium in size.
Hierarchical Method	- Hierarchical method is also known as connectivity based clustering.
Grid-Based Method	- Multi resolution grid structure approach. - Examples STING (statistical information grid) and CLIQUE (clustering in Quest) [2].

Table 1.1 Representation of cluster methods

1.1.1.3 REQUIREMENTS FOR CLUSTER ANALYSIS

1) **Scalability**: - clustering algorithms should be highly scalable because small data sets having few data objects clustering algorithms provide good result but in the

case of web scenario which have the large data sets having millions or billions data objects clustering algorithms provide biased result.

2) **Ability to deal with noisy data**: - In the real world there are many data sets which contain missing values, erroneous input for example sensor reading. In the sensor reading, readings become inaccurate in value due to the sensor mechanism or by transient objects from surrounding. So due the presence of error outlier in the data sets clustering algorithms produce the bad quality cluster. That why we need such clustering algorithms or methods which can deal with noise present in the data sets.
3) **Ability to deal with different type of attribute**: - Now a day applications require clustering techniques should be deal with complex data types such sequence, image and graph. Many clustering techniques or algorithms are there which deal with only cluster numeric data.

1.1.1.4 APPLICATION OF CLUSTER ANALYSIS

There are many fields where cluster technique can be applied: -

1) **Image segmentation**: In the field of image segmentation, clustering technique is used to divide the digital image into distinct region.
2) **Market research**: With help of clustering technique, help in finding the product development, customer behavior and market trends etc
3) **Educational data mining**: In this field cluster analysis used to find those colleges and universities which resemble to the similar property.
4) **Crime analysis**: In the crime analysis, cluster technique is used to identify which area having more crime.
5) **Grouping of shopping items**: With the help of cluster analysis, all the shopping items are group into unique set of products on the web for example eBay which is not using the concept of SKU.

There are two parameters through which we can measure cluster quality: -

1) Intra-cluster: Intra-cluster similarity percentage should be high for the good clusters.
2) Inter-cluster: Inter-cluster similarity percentage should be low for the good clusters.

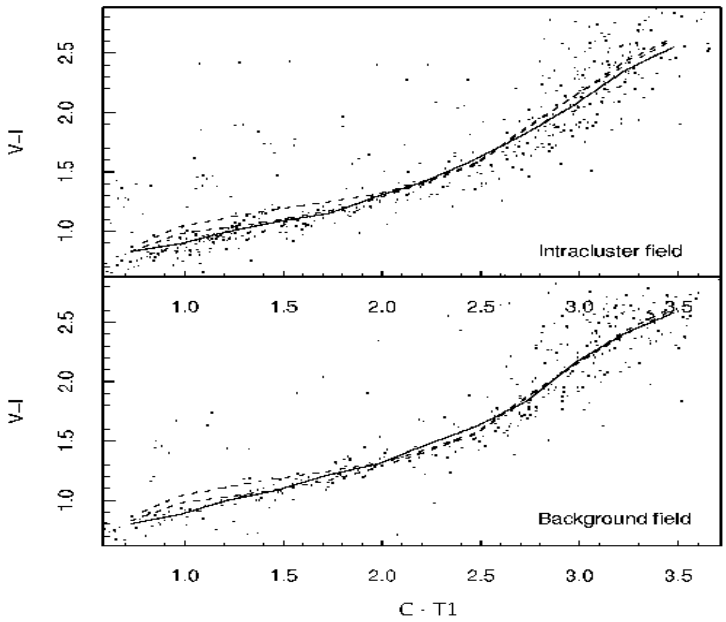

Fig1.4 Representation of Intra cluster

1.2 TEXT MINING

Text data mining or knowledge discovery in data base is also known as "Text Mining". In the text mining new knowledge can be discovered by the analysis of large text. Unstructured database is used in the text mining and the search mode is Opportunistic.

Fig1.5 Text mining

1.2.1 APPLICATIONS OF TEXT MINING

1) **Industry: -** In industry, text mining helps in finding competitor products , new products in the market, strategy of competitor from the web pages.
2) **Job seekers: -** With the help of text mining, parameters can be indentified for the fresher's, unemployed or job seekers. For example www.fresherworldjob.com, www.naukari.com.

1.2.2 TEXT MINING CONSIST TWO PHASES

1) **Text refining**: - In this phase text document will transform into the intermediate form
2) **Knowledge distillation:** - In this phase pattern will be deduce from the intermediate form.

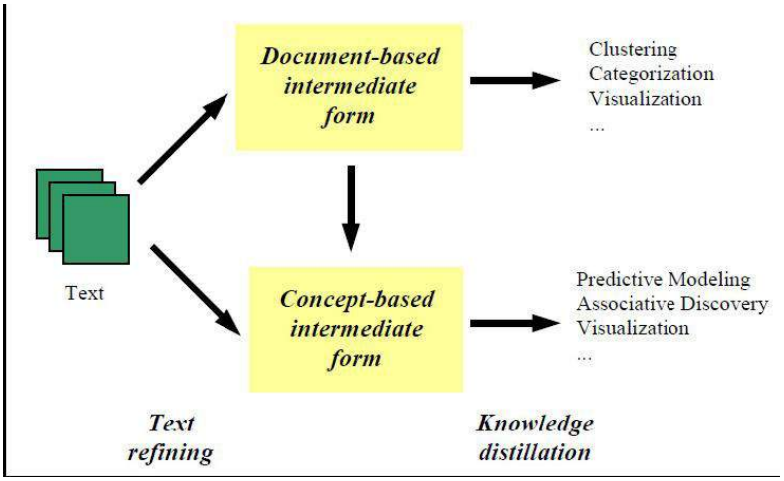

Fig1.6 Text Mining Phases

1.2.3 TEXT MINING METHODS

There are two methods through which we can mine the text:

1) **Information Retrieval**: - Information Retrieval is a technique through which user finds the useful information from the large document text. There are four phases in information retrieval: -
   a) **Indexing**: - In this phase, raw documents convert into the document representative which can be useful by information retrieval software [2].
   b) **Query formulation:** - In this phase user will enter query for the retrieval of information from the large document text [1].
   c) **Comparison:** - In this phase, query enter by the user will match with the stored document and then on the basis of matching query result will be formed [1].
   d) **Feedback**: - In this phase, the user may simply communicate to the information retrieval system judgements about the desireablity of each

retrieved document, and the system may implictly updatethe query itself. This latter process is refred as relevance feedback [2].

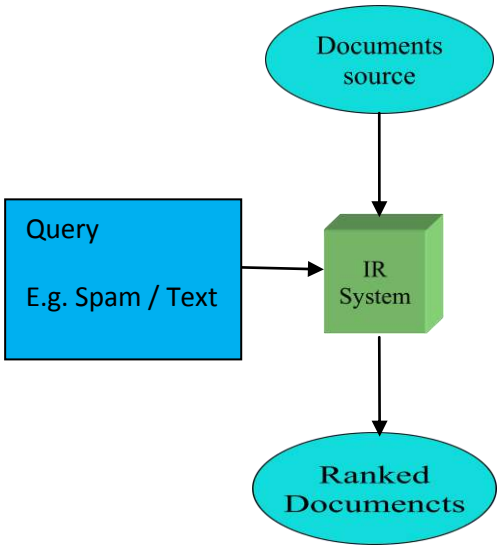

Fig 1.7 Process of Information Retrieval

2) **Clustering: -** In this method, similar text is grouped together in one cluster and dissimilar text is grouped in another cluster. There are various methods in cluster analysis through which we can group the similar text in one cluster. The main objective of clustering is to find the correct sets of documents.

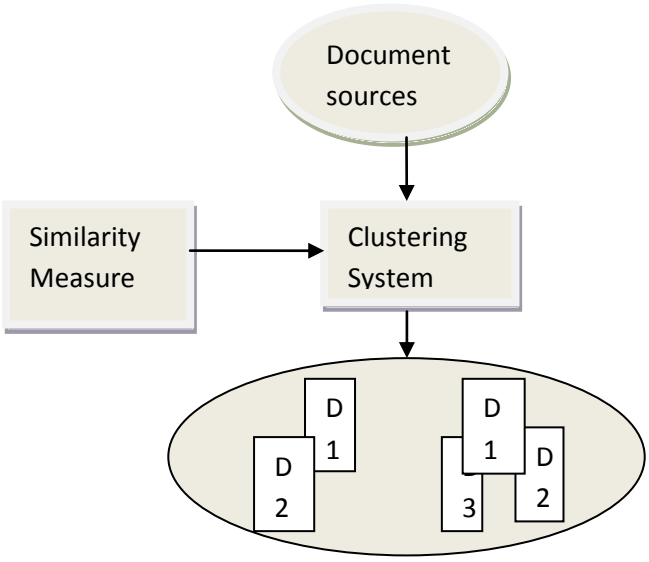

Fig1.8 Cluster finding common words in documents

1.2.4 CHALLENGES IN TEXT MINING

There are various challenges faced during text mining given below:

1) There can be complex relationship between concepts in the text mining.
2) In the text mining, records are not structurally identical as done in the data mining.
3) There can be ambiguity and context sensitivity in the text mining. For example: apple can be fruit name or it can be company name.

CHAPTER 2
REVIEW OF LITERATURE

Samiran Ghosh et.al (2011), ETL is essential part for the execution of data warehousing. There are the two major facets of this paper first that, it represents the ETL log and second, it apply the outlier detection technique on the same ETL logs. In this paper they have used the outlier dictation technique to find the processes, which is unreliable as of group. They captured the lot of execution trace in term of log files with the plethora of information. If there is outlier in the production process then it will be act as signal otherwise it will consider as noise. In the extract phase they use the text parsing for extracting the feature from log files. Consecutively they applied the outlier detection technique on the logs as a result they reduced the domain of detailed analysis 500 logs to 44 logs in term of percentage 8% [11]. By applying this approach, organization will not require any important investment because cost of text parsing application and data mining application is reliable and low.

Wafa. Maitah et.al (2013), The Information retrieval technique is used to discover the useful information from the large document text to the users. Useful Information can be retrieved within the same document or it can be retrieved from the different documents. In this paper they have applied the adaptive genetic algorithm in information retrieval and tried for the optimized queries. Queries will call to be good if they state the precise information to the users and also reduce the number of iteration. But to optimize the query is not the very easy task, so more efforts had been applied to improve the query. In this paper data set has been taken from the proceeding Saudi Arabian national conference which is 242 Arabic's abstract [17]. After that preprocessing phase will start in which all the data will extract from the each documents and elimination stop- words and stemming will also be apply. When the terms have been assigned to the documents, then document weight will be determine with the help of cosine similarity. In this research they mentioned, the traditional similarity approach i.e. Vector space Boolean [17].They chooses the AGA as compare to GA because of the following reasons: -

1) AGA provides the better performance when we are using the crossover and mutation operators with the variable probabilities in opposite GA will provide the

fix values during the execution and also remain unchanged for these above mentioned operators.

2) AGA will provide the adaptive adjustments of crossover and mutation probabilities which will lead to achieve the better performance then GA [17].

Rupali Gill and Jaiteg Singh (2014), Data quality is the major issue in the data ware house. Good data quality is considered to be following attribute correctness, accuracy, consistency, and timelines. Good data is very demanding because every organization tries to improve its services for the end user or customers. The majority of the research has been applied on the data quality in data ware house ETL process .ETL is considered to be high-quality tools to integrate the data from the heterogeneous but to handle the ETL process is not very straightforward and easy task to handle. Before the data to be load into data ware house it should be clean because the data is integrate from the different sources. In this paper researchers had been proposed the framework which is using the good quality of data in ETL in data ware house. They take three different data sources for extracting the data i.e. My Sql, Ms excel Ms Access [18]. After that cleansing and transforming step is start in which set of rules is used to transform the data to the destination. In this step, they tried to resolve name conflict structure conflict, handling missing values and applied domain constrain check. Loading is the final step in which the data is load into the destination place. Loading phase will include both fact table and dimensional tables.

Mong Li Lee et.al (2002), in this research work researchers had implemented the novel integration strategy which includes the clustering of DTD (document type definition) of XML data sources [3]. The main motive of this research is to contribute towards the integration of real world DTD to express the XClust approach. In this experiment, they make the cluster of DTD which are similar in both the way semantically and structurally. With the help of cluster conciliate the resembling DTD is simple as compare to conciliate the DTD that are different in semantic and structure. There are two main phases of this paper in the first phase they determining the degree of similarity of DTD's. They compared the DTD elements not only considering the linguistic and structural information of DTD element but also examine the context of DTD elements. The similarity between the DTD elements depends upon the semantics, immediate descendents and leaf context. In the second phase they integrate the real world DTD. In

this work, researchers had explained that how integration of DTD's and leaf context information plays a very critical role for matching DTD elements correctly [3].

Varun Kumar and Nisha Rathee (2011), in this paper researcher have proposed the integration process on cluster and classification technique in data mining. To implement this experiment, they take the dataset from "Fisher Irish dataset" in which it contains 5 attributes and150 instances [12]. In this research, they compared the simple classification technique in data mining with the integration of clustering and classification technique. To make this experiment prosperous, they make use of J48 classifier of simple classification technique with the result of integration of clustering and classification technique by using tool WEKA. By comparing these, researchers had demonstrated that by integrate the clustering and classification technique gives higher outcome than the simple classification technique. Observation of this research shows the following results: Error Rate: by comparing j48 with the K Mean cluster is 0.0909, Without K Mean cluster, isolate classification is 0.1713. Accuracy: J48 with K Mean cluster is 98.66%. Sensitivity: J48 with K Mean cluster is 1, Without K Mean cluster, isolate classification is 0.98 [12].

Marcello Leida et al. (2013), in this paper, they have explained the mapping model, which also refer as an ontology, which will solve the problem of semantic heterogeneity. In the data integration step they integrate the data from the different database and make them as a unified view which can give useful information to the users. However, the problem with this is that we are not able to integrate the data properly from different databases. In the previous studies, researchers have applied two types of approaches i.e. declarative and procedural. These approaches support the coupling of the data, which depend upon the pre-defined queries. However, the problem in these approaches that they are very time consuming and found very costly during implementation. These approaches are not very flexible with the modern tools .After that researchers had applied the mapping model approach, which will integrate the data from different data sources database. This paper has also introduced the concept of semantic identifier and the semantic join. Semantic join provide the common set of information to the set of query, which will support the mapping language technique to integrate the data from different set of database. Semantic identifier is the general approach to the entity resolution. After defining these attribute they fuse these element together to represent the same entity result. The result has submitted to the system after the fusion step and with the help of

which mapping result generates. Ontology virtual-A box has used to represent the mapping model results. In this paper, mapping model has been proposed because of the following benefits:

1) It saves the cost of the implementation as well as flexible in implementation of mapping model.
2) Mapping model is able to answers the complex queries.
3) Mapping model is different from the above-explained traditional approaches because this model saves the implementation time.

Shaowei Wang et al. (2012), in this paper researcher have proposed the similarity metrics to infer semantically related terms. To implement this experiment researchers have used the concept of documents tags which is use to describe the important feature of software products. Tag concepts are being use in this research because it relates the term in semantic manner which is present in different documents. For this, researchers have collected the ten thousands of projects and collection of their tags from the free code dataset. K Medoids algorithm had been used in this paper. Approach used in this paper consist two steps, in the first step they calculated the similarity between each pair of terms. In this step similarity metric has been proposed based on the documents in which tag is used by the terms. In the second step which is based on the similarity metrics, they apply the K Medoids algorithm in which they infer the taxonomy of terms.

Huong Morris et al. (2008), in this paper, researchers have implemented the Callisto architecture with the help of which ETL tool becomes more efficient to integrate the business objects [9]. To implement this experiment, researchers have used the Callisto architecture because this architecture provides ability to examine the WPC catalog. In Callisto to model the ETL process, they have used the RDA (rational data architect) and rational rose to model WPC objects. They have developed the four types of operators, which represent the business objects for import and export function in WPC. These operators are java-based, through which business objects can be integrated, assembled or disassembled. Approach followed in Callisto, that to generate the java code they captured the key constraints between objects by UML and EMF modeling [9]. This experiment helps in accessing the product data management. For the future scope, researchers can develop more operators, so that master data management system can also be validating by using this approach. In the field of business, business objects play important roles

because these objects are responsible for capturing the business concepts in semantic manner which directly impact to business processing.

Soumi Ghosh and Sanjay Kumar Dubey (2013), in this paper researchers have compared the two clustering algorithms first is centroid based K-means and second one representative object based Fuzzy C-Means in data mining technology. Tool used for implementing this comparison is MATLAB. To test the efficiency of these algorithms data sets has taken from UCI machine learning repository and iris plant data set. In the iris plant data set, there are total number of attribute is 5 out which 4 are numeric i.e. sepal length, sepal width, petal length and petal width and one is non numeric [15]. In this experiment, researchers have proved that efficiency of K mean algorithm is more accurate than fuzzy c mean. K-means provide the better result than fuzzy c-means. Computational time of fuzzy c-mean algorithm is more than k means algorithm. Time complexity of K Means algorithms is o(ncdi) and a fuzzy c-means algorithm is $o(nd^2 ci)$ [15].

Hao et al. (2008), the researchers has explained that Data integration is very big issue because it combines the data from different databases and explains how to represent the accurate information to the users. Many problems had faced during data integration:-

1) **System Heterogeneity**: - This type of heterogeneity occurs due to the incompatibility with hardware, operating and networking.
2) **Global Model**: - How to build the global model for the users so that, they can view all the information in one model without any difficulty.
3) **Query processing**: - How to process the query in effective manners because quality of the good information depends upon query execution.
4) **Semantic Heterogeneity**: - Semantic heterogeneity occurs when the dataset for the same domain have been created independently by the different parties.

Searls et al. (2005), in this paper, they mentioned that integration is the process through which we collect the data from the different sources and combine them to make unique view of the information to the users. There are many applications of data integration, which have played a vital role. In the area of the business intelligence in which they are not only collecting the information about the employees of the company but also kept the information about the individual customers, what should be the current sales to improve the real fact about the market? Major problem faced in the integration is in the area of

heterogeneity. This is the main problem because data are not easily interoperating with the other databases and different parties independently create the dataset for the same domain. They also explained some of the approaches in the integration. First approach is point-to-point model in which application can correspond directly with each other. However, this approach is useful in the small applications because there is limited number of sources, which can interact each other. Second approach is integrated hub model as the name signifies that hub has placed between the applications so that each application can communicate with that hub rather than each other. This is applicable for the large application. This paper has also explained about the ETL tool. ETL refer to that process in which firstly we extract the data from the heterogeneous data sources and then transform that data into accurate format and after that, we load that data into destination.

Liu et al. (2010), they have explained the Decision support system, which is well-established research area. The researchers had explained the current position of the IDSS key integration issues and the positive impact on the DSS. The researchers had reviewed the papers that had concentrated about the technologies and methods used for the integration in decision support systems. The researchers had explained that integration is the major problem in DSS. For modeling the decision support system to gives high performance more efforts has been paid for the integration the data in qualitatively. More challenging part in the integration is to provide the data in flexibility and reliability manner. The researchers had explained integration as future challenge in DSS performance. They mentioned that more support will be provided by DSS if there is better integration of data.

Bellatreche et al. (2013), they explained the aspire of data integration in which we combine the data from different source of databases and make unified view for the users
Integration system has based on the three architectures:-
(a) **Materialized**: - It is the repository for the data duplication. The best example for the materialized is data warehouse, which is related to the business intelligence.
(b)**Virtual**: - Virtual is the logical view of the sources. Mediator has explained in the virtual in queries of the users are placed into the data source.
(c)**Hybrid**: - The Hybrid is the combination of both the source and global schema.

Alon Halevy (2005), in this paper, the researcher has mentioned about the data integration where the data is residing in different sources and should be combine together to give the useful information to the users. They explained about the semantic heterogeneity in which datasets for the same domain has created by independent parties by which this problem occur. This problem can be also seeing in XML documents. In the real world information from the enterprises faces many problem when they accessing and analyzing the data which is residing in different sources. For example in the enterprises, if they want to view the customer's details as a single view they face the problem because they have to collect the data from the multiple databases. To maintain such large data is just the infancy for the enterprises. The problem is that we are extracting the semi-structured data from the unstructured data .Therefore we should maintain the data values and attribute properly.

Hakimpour et al. (2005), in this research, researchers have explained the problem during handling semantic heterogeneity in data integration. Semantic heterogeneity refers to different meaning in the data [6]. In this paper researchers have tried to represent the global schemas. To represent the global schema, they use the concept of formal ontology which resolves the problem of heterogeneity. Researchers use the concept of ontology in data integration for the following reasons:

1) Completeness: - If the number of non-overlapping similarity relations detected is small, our approach will result in low-quality integration, as it will rather be a union of schema definitions than a true integration [6].
2) This approach provides the schema definition to the community formal ontology.

In this research, constraints to the schema definitions also used which makes this research successfully.

Andreas ET al. (2003), in this paper researcher had implemented the semantic metadata by using the generating RDF semantic web schema. In this paper researchers approach is different from the previous approach with the respect of both structural information as well as content information which is analyzed by the previous knowledge. The current schema that researchers applied is able to either linguistic or semantic annotation of data pieces in the web documents using prior knowledge or natural language processing (NLP) [4]. In this research schema that have been generate from the documents is in the form of structured way which can be easily implemented. This research also helps in solving the

mapping handling heterogeneity i.e. helps in mapping between the structured data sources and semantic web schema where as previous approach was not feasible in both the cases semi structured as well unstructured. To make this research purpose full, researchers had extracted the information from the web documents. Researchers have followed the following step while doing this research:

1) In this step identification of the object resources with the help of hyper link and URL.
2) In this step, metadata of object resources are extract as the attribute of the object resources. In this it will include name of the author, size of the file or content information etc.
3) In this step documents pre-processing will be applied. In this unstructured information must be pre-processed so that document can be handled automatically. In this weight age approach had also been used.

CHAPTER 3
PRESENT WORK

3.1 PROBLEM FORMULATION

In the data integration process, data combine from the multiple sources but ignore the similarity between texts.

Clustering method is an indirectly approach for documents or texts similarity, this approach provide semantic similarity of the texts without increase the complexity of process.

In the previous research work, for documents similarity using the concept of Euclidean which is only consider about geometrical distance. So it increase error or reduce the accuracy by making the faster cluster.

In the previous research work, algorithms K-Mean and UPGMA implemented individually, so they cannot take advantage of both approach in one algorithm.

In our research work, we integrate the texts with fewer features by using vector space model (text frequency-inverse document frequency TF-IDF) and apply K-Mean with cosine similarity because cosine similarity represents actual similarity and difference between documents. We have also take advantage of hybrid model of K-Mean and UPGMA with cosine similarity and enhance the capability K-Mean clustering algorithm because it will not take central randomly**.**

3.2 OBJECTIVES

1) To build dataset manually for our experimental works.
2) To implement the K-Mean and UPGMA with Euclidean distance clustering methods.
3) To implement the vector space model on different documents.
4) To implement K-Mean with cosine similarity.
5) To build and implement a Hybrid model of K-Mean and UPGMA with cosine similarity.
6) Calculate the accuracy, precision, recall and f-score of all implemented algorithms and choose the best method for integrating texts.

3.3 RESEARCH METHODLOGY

Research methodology is method through we research the problem in the systematically and the scientific way. For the research methodology, we should understand the methodology to be chosen for the research. In the research methodology there are various steps are included, through which we can research our problem.

In our research work we followed steps: -

1) In the first step we will extract the texts from the different sources.
2) After extracting the text, pre-processing phase will start in which tokenization and stop world removal will be applied on the text.
3) In the third step, after pre-processing phase we will calculate the TD-IDF phase in which document weight age will given to texts and represent in the vector form.
4) After completion these steps, K-Mean and Hybrid approach UPGMA and K-Mean cluster algorithm with cosine similarity will apply in which there will be automate the integration of data will generate which is coming from the different sources.
5) In the last step we will calculate the accuracy, precision, recall and f-score of K-Mean and Hybrid approach.

RESEARCH METHODLOGY

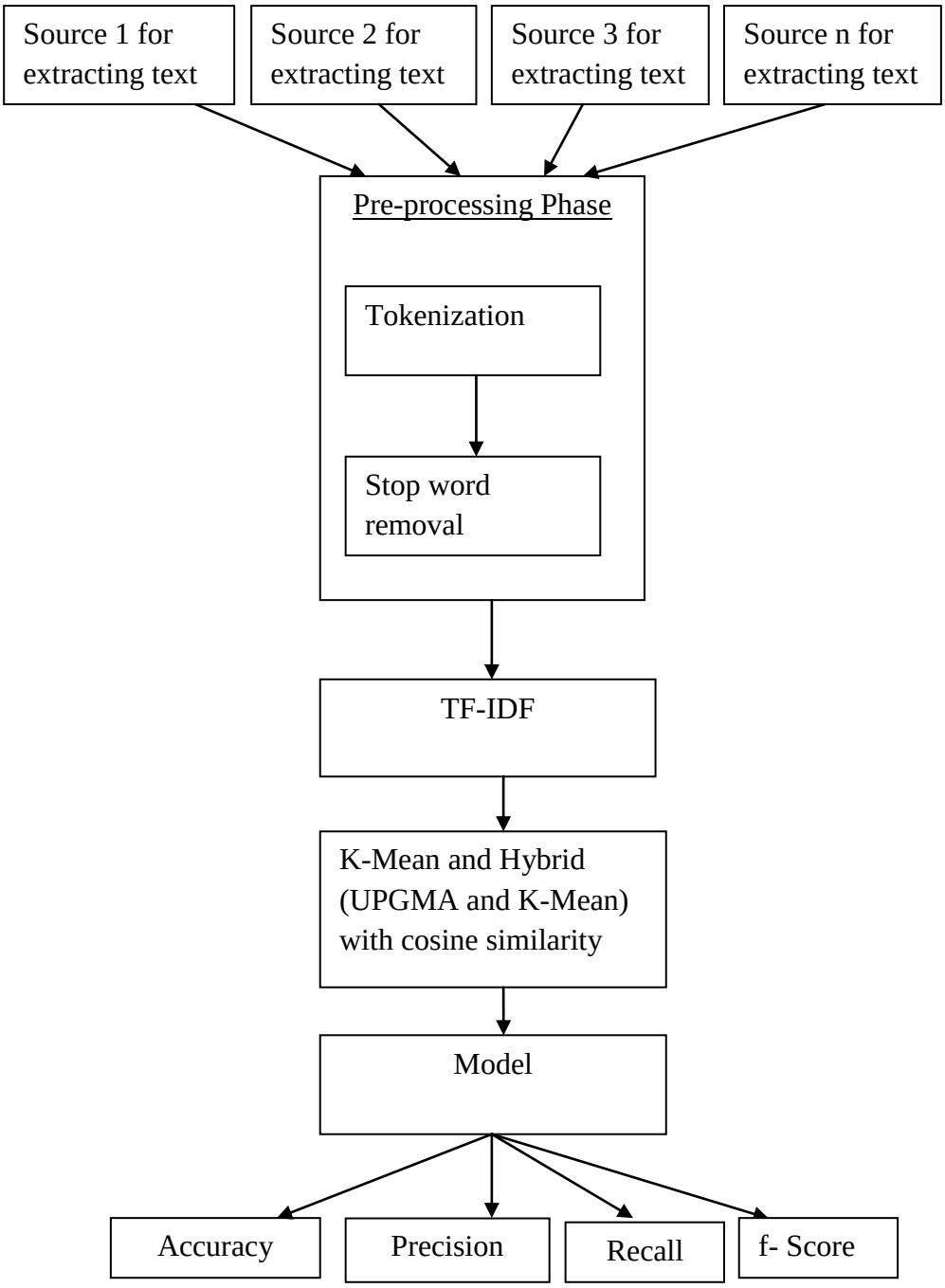

Fig 1.9 Research methodology

Tokenization

In the process of tokenization, text streams are broken into some meaningful elements, called tokens. These tokens become the input to the text mining or parsing for further processing. There are many obstacles while tokenization like white space and punctuation

may or may not be included in the list of tokens. Tokenization is difficult to apply in Scriptio continua languages in which there exist no word boundaries like Thai. Software used for tokenization is apache openNLP and U-Tokenizer [2].

Cosine similarity

Cosine similarity is used to measure similarity between two vectors of inner products that measure the cosine of the angle between them [1]. The value for 0^o cosine is 1 and for other angles its will be less than 1. The Cosine Similarity is mainly used in higher-dimensional vectors spaces. In the field of data mining, cosine similarity measure the cohesion within the clusters.

CHAPTER 4
RESULT AND DISCUSSION

TOOLS

1) **Python**: - Python is a multi-purpose programming language helps in creating web development, game development and efficiently text processing. It is not case insensitive language i.e. Apple and apple are two different identifier in python. In python, there are various standard libraries that support internet protocols such JSON, HTML, XML and Email Processing.
2) **NLTK**: - Natural Language ToolKit. NLTK is free and open source ToolKit which provide (implement) platform for python language. NLTK 3 version is used for this experiment. Python version should be 2.6, 2.7 or 3.2 + for NLTK.
3) **Sk learn**: - It is scipy Toolkit. Scikit-learn is an open source machine learning library for the Python programming language [20]. Scikit-Learn tools used for efficient data mining and data analysis. We have used 0.16.1 scikit-learn version [20].

We have measured our experimental results on the basis of four metrics

1) ACCURACY

Accuracy = $\frac{TP+TN}{P+N}$ [1]

Whereas TP = True Positive

TN = True negative

P = Number of positive tuples

N = Number of negative tuples

Sensitivity = $\frac{TP}{P}$ and Specificity = $\frac{TN}{N}$

2) PRECISION

Precision = $\frac{TP}{TP+FP}$ [1]

Precision is defined as closeness of two or more measurements to each other [10].

RECALL

Recall = $\frac{TP}{P}$ [1]

Recall can be measure of exactness [10]. Recall is similar as sensitivity [1].

3) F-SCORE

F-Score = $\frac{2\times Precision\times recall}{Precision+recall}$ [1]

EXPERIMENTS

K-Mean with Euclidean distance

In the graph of K-Mean with Euclidean distance show that accuracy of first two clusters remain increases but at the fourth cluster suddenly decreases because K-Mean with Euclidean distance unable to detect the document similarity in correct way.

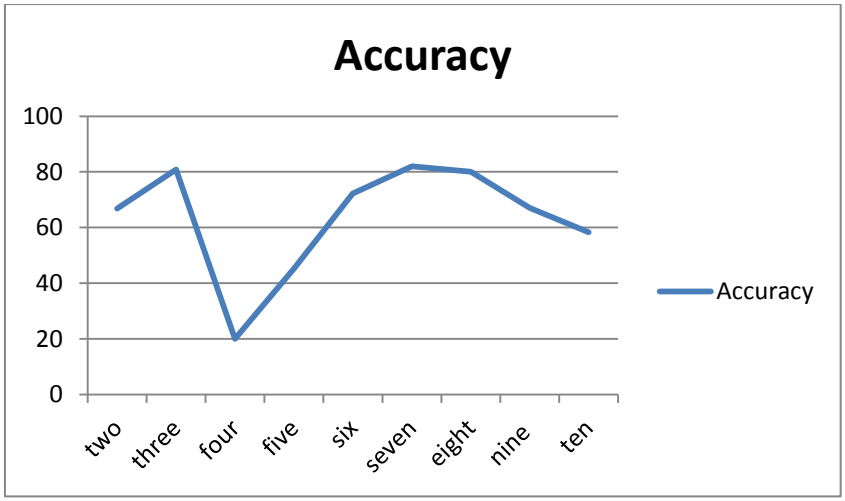

Fig 4.1 Accuracy of K-Mean with Euclidean distance

Cluster	Accuracy
Two	66.8
Three	80.9
Four	20
Five	45.2
Six	72.2
Seven	82
Eight	80
Nine	67
Ten	58.28

Table 4.1 Representing K-Mean Accuracy values

In the graph of K-Mean with Euclidean distance representing precision, recall and f-score below shows that percentage of these metrics first increase and then decreases in invariant manner because K-Mean Euclidean with distance unable to detect document similarity due to which of these metrics decreases.

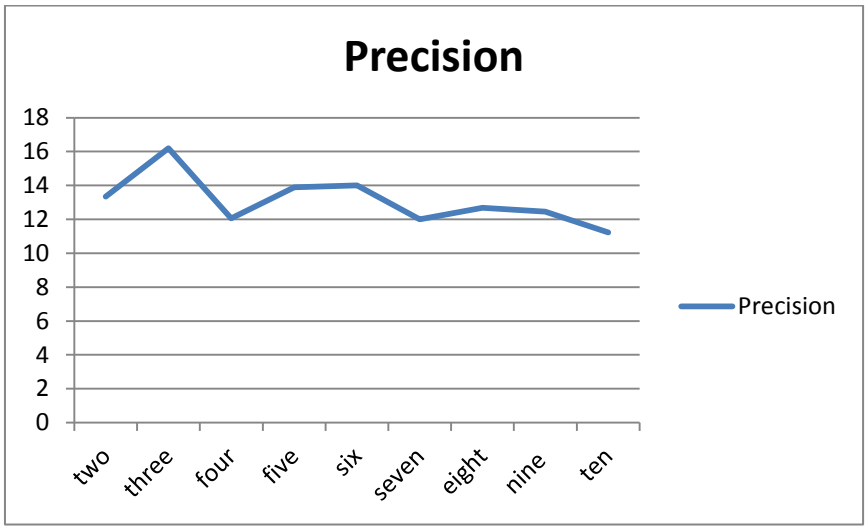

Fig 4.2 Precision of K-Mean with Euclidean distance

Cluster	Precision
Two	13.34
Three	16.2
Four	12.06
Five	13.9
Six	14
Seven	12
Eight	12.678
Nine	12.45
Ten	11.222

Table 4.2 Representing K-Mean precision values

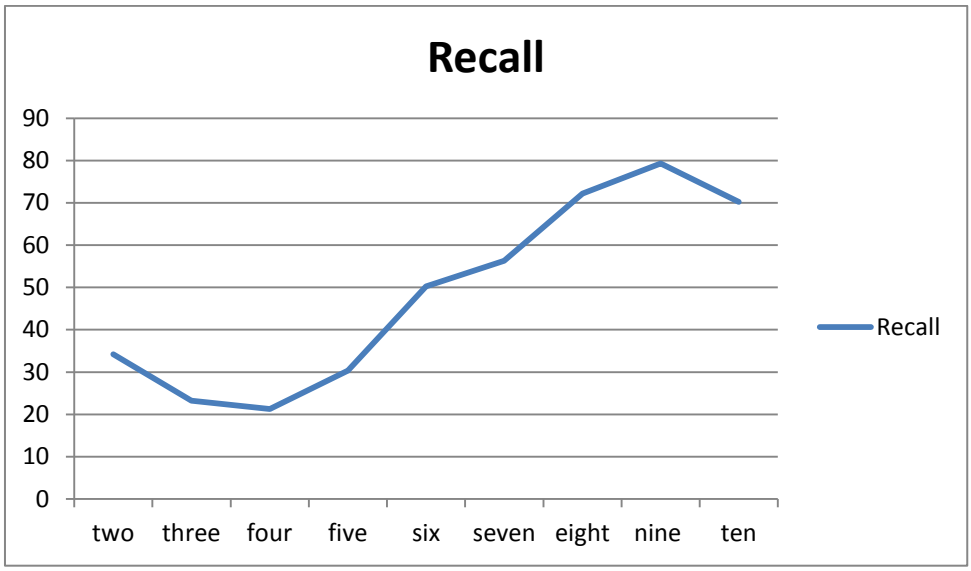

Fig 4.3 Recall of K-mean with Euclidean distance

Cluster	Recall
Two	34.23
Three	23.23
Four	21.23
Five	30.34
Six	50.22
Seven	56.28
Eight	76.22
Nine	79.33
Ten	70.234

Table 4.3 Representing K-Mean Recall value

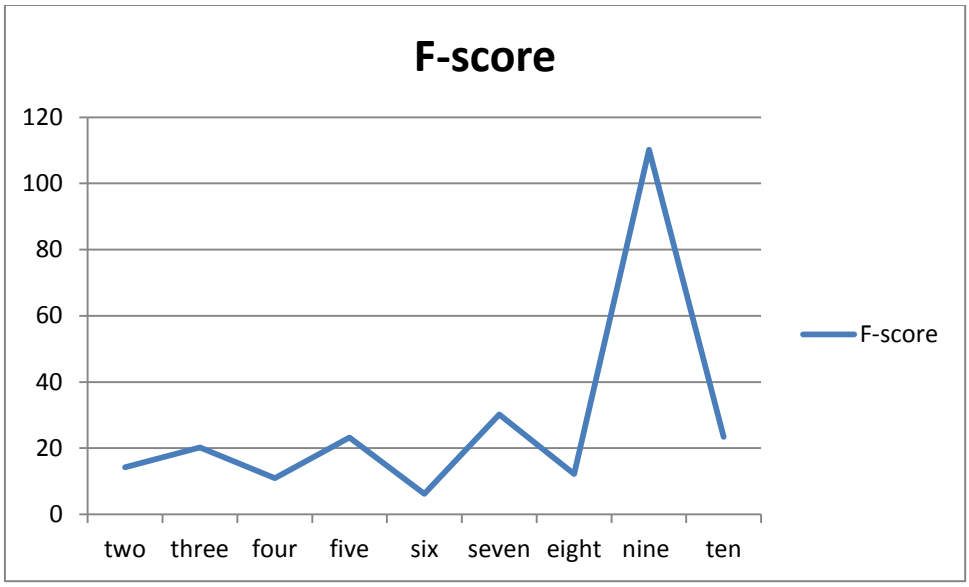

Fig 4.4 F-score of K-Mean with Euclidean distance

Cluster	F-score
Two	14.23
Three	20.23
Four	11
Five	23.23
Six	6.23
Seven	30.23
Eight	12.23
Nine	110.23
Ten	23.45

Table 4.4 Representing K-Mean F-score value

UPGMA with Euclidean distance

In the graph of UPGMA with Euclidean distance representing accuracy, precision, recall and f-score shows that percentage of these metrics first increase and then decreases in invariant manner because UPGMA with Euclidean distance unable to detect document similarity in correct order due to which error occurs and results of these metrics decreases.

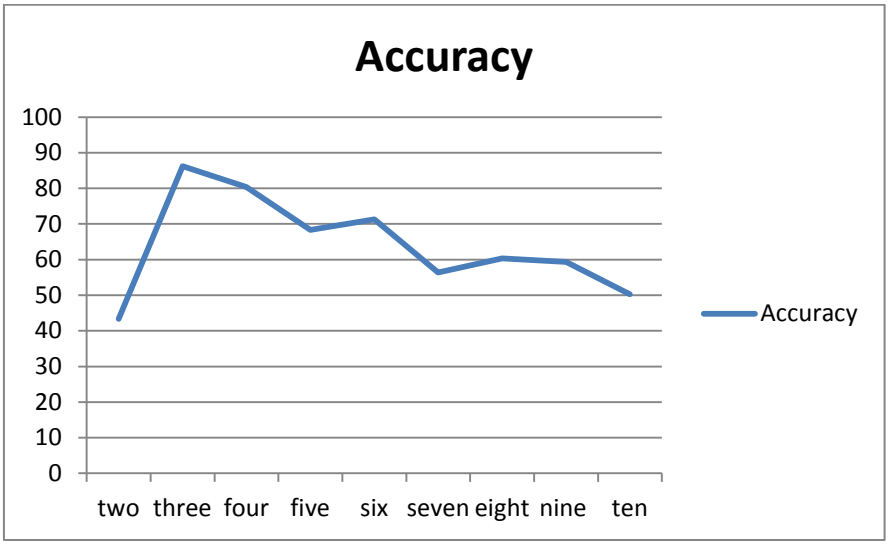

Fig 4.5 Accuracy of UPGMA with Euclidean distance

Cluster	Accuracy
Two	43.34
Three	86.23
Four	80.34
Five	68.34
Six	71.3
Seven	56.34
Eight	60.34
Nine	59.34
Ten	50.3

Table 4.5 Representing UPGMA Accuracy value

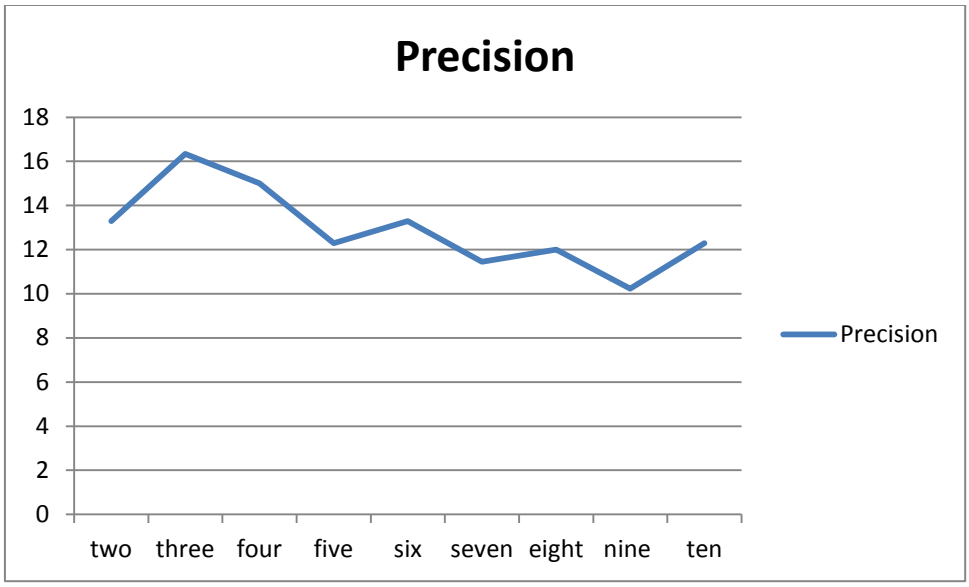

Fig 4.6 Precision of UPGMA with Euclidean Distance

Cluster	Precision
Two	13.3
Three	16.34
Four	15
Five	12.3
Six	13.3
Seven	11.45
Eight	12
Nine	10.23
Ten	12.3

Table 4.6 Representing UPGMA Precision value

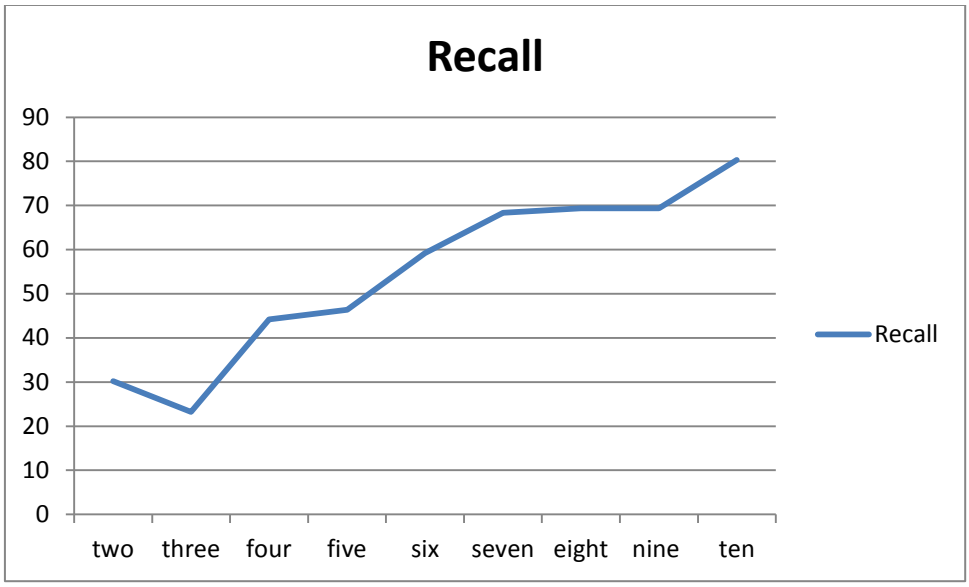

Fig 4.7 Recall of UPGMA with Euclidean Distance

Cluster	Recall
Two	30.23
Three	23.23
Four	44.23
Five	46.34
Six	59.23
Seven	68.34
Eight	69.34
Nine	69.3453
Ten	80.31

Table 4.7 Representing UPGMA Recall value

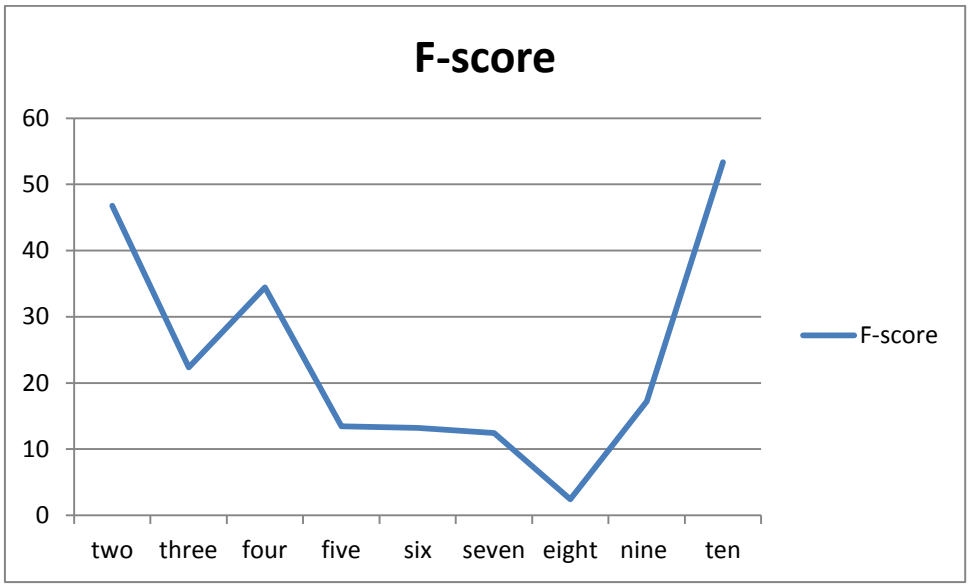

Fig 4.8 F-score of UPGMA with Euclidean Distance

Cluster	f-score
Two	46.78
Three	22.34
Four	34.45
Five	13.45
Six	13.23
Seven	12.45
Eight	2.45
Nine	17.23
Ten	53.34

Table 4.8 Representing UPGMA F-score value

K-Mean with Cosine Similarity

In the graph of K-Mean with cosine similarity representing accuracy, precision, recall and f-score shows that percentage of these metrics increases or decreases but with slightly difference between them. The performance of these metrics is better in K-Mean with cosine similarity because cosine similarity represents actual similarity between documents.

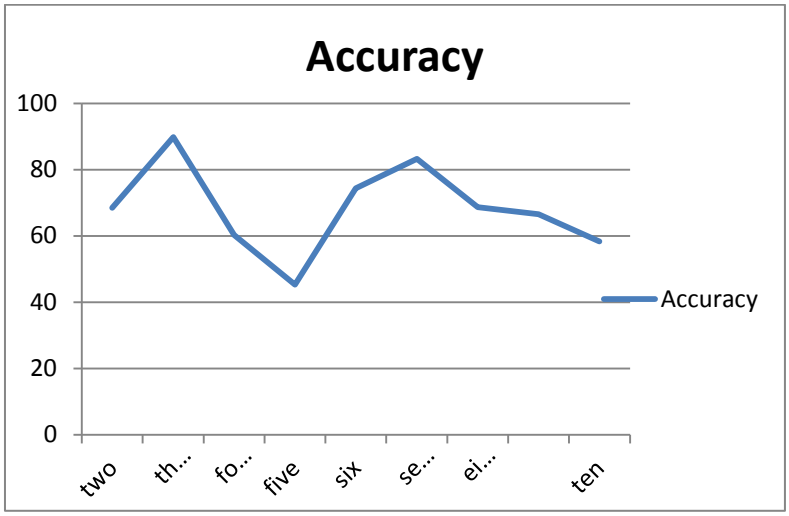

Fig 4.9 Accuracy of K-Mean with Cosine Similarity

Cluster	Accuracy
Two	68.5
Three	89.9
Four	60.29
Five	45.28
Six	74.4
Seven	83.3
Eight	68.7
Nine	66.6
Ten	58.33

Table 4.9 Representing K-Mean with cosine similarity Accuracy value

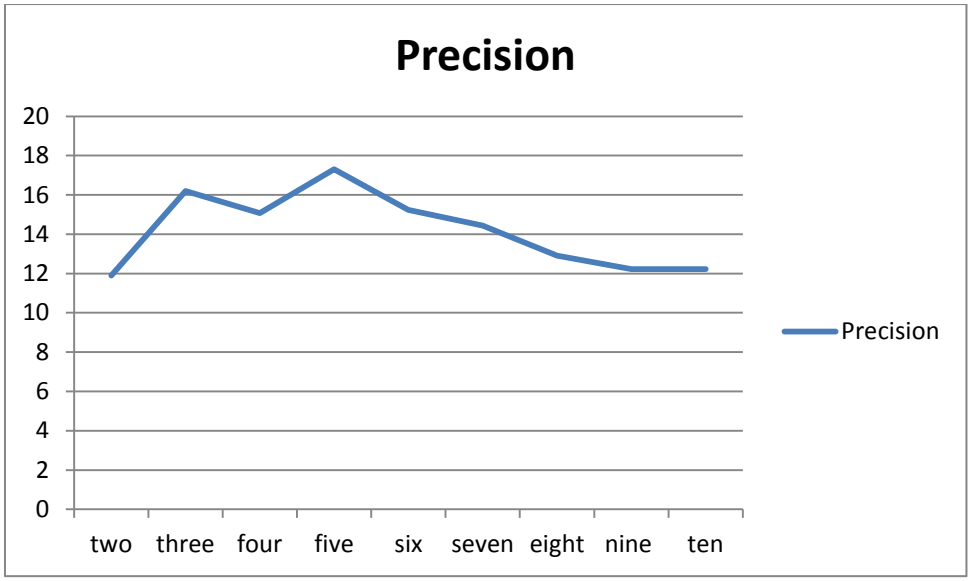

Fig 4.10 Precision of K-Mean with Cosine Similarity

Cluster	Precision
Two	11.9
Three	16.2
Four	15.07
Five	17.3
Six	15.23
Seven	14.44
Eight	12.9
Nine	12.22
Ten	12.22

Table 4.10 Representing K-Mean with Cosine Similarity Precision values

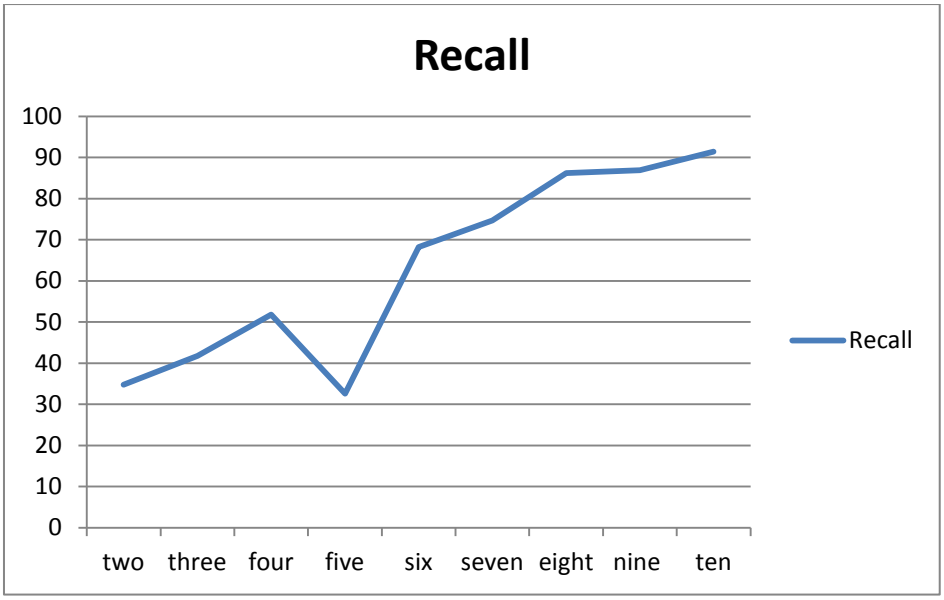

Fig 4.11 Recall of K-Mean Cosine Similarity

Cluster	Recall
Two	34.8
Three	41.8
Four	51.8
Five	32.6
Six	68.2
Seven	74.7
Eight	86.2
Nine	86.92
Ten	91.45

Table 4.11 Representing K-Mean with Cosine Similarity Recall value

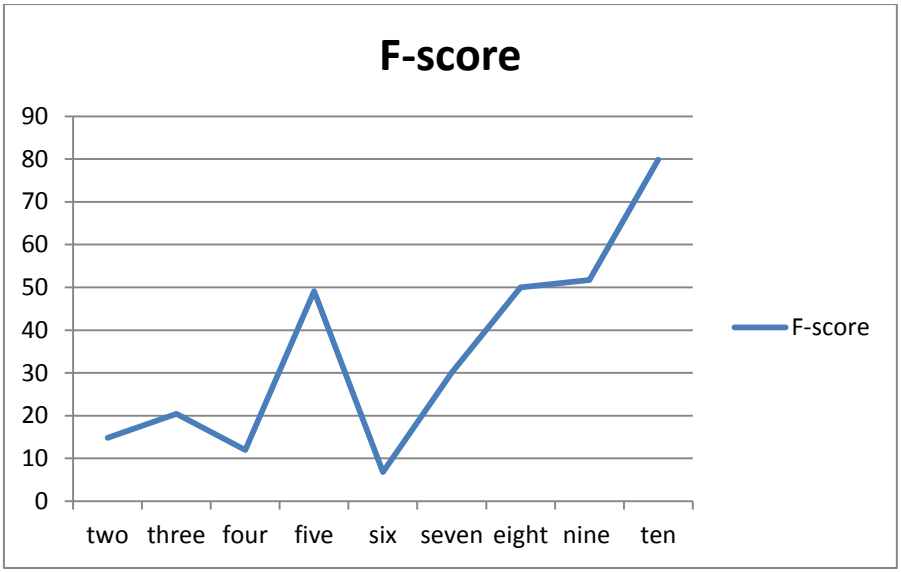

Fig 4.12 F-score of K-Mean with Cosine Similarity

Cluster	F-score
Two	14.8
Three	20.4
Four	11.98
Five	49.1
Six	6.8
Seven	30.1
Eight	50
Nine	51.7
Ten	79.9

Table 4.12 Representing K-Mean with Cosine Similarity F-score values

Hybrid UPGMA and K-Mean

In the graph of Hybrid approach UPGMA and K-Mean with cosine similarity representing accuracy, precision, recall and f-score shows that percentage of these metrics increases or decreases but with slightly difference between them. The performance of these metrics is better in hybrid UPGMA and K-Mean with cosine similarity because cosine similarity represents actual similarity between documents.

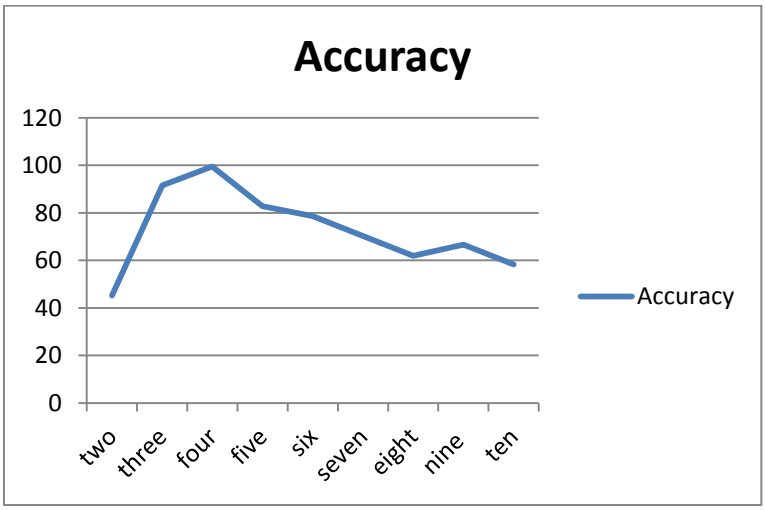

Fig 4.13 Accuracy of Hybrid UPGMA and K-Mean

Cluster	Accuracy
Two	45.2
Three	91.5852083
Four	99.4486198
Five	82.7819531
Six	78.6152864
Seven	70.2819531
Eight	61.9486198
Nine	66.6666667
Ten	58.3333333

Table 4.13 Representing Hybrid with Cosine Similarity Accuracy values

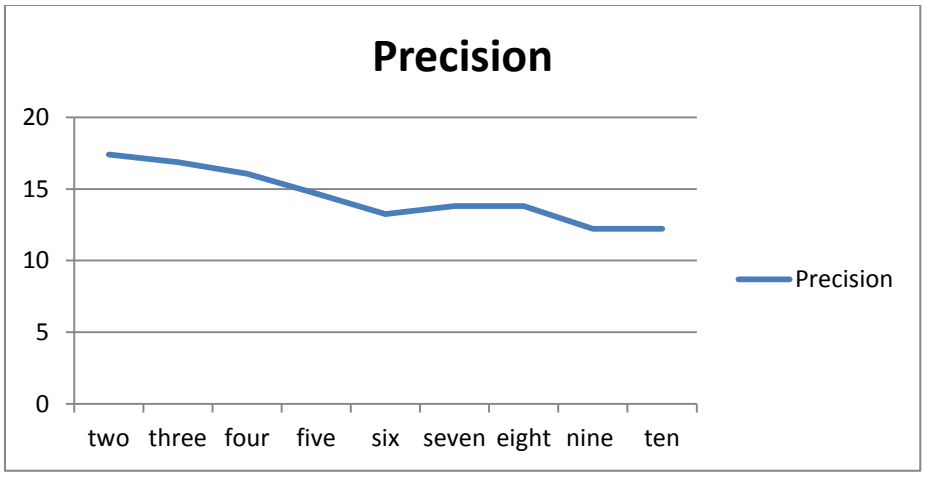

Fig 4.14 Precision of Hybrid UPGMA and K-Mean

Cluster	Precision
Two	17.3992674
Three	16.8650794
Four	16.0714286
Five	14.6825397
Six	13.2539683
Seven	13.8095238
Eight	13.8095238
Nine	12.2222222
Ten	12.2222222

Table 4.14 Representing Hybrid with Cosine Similarity Precision values

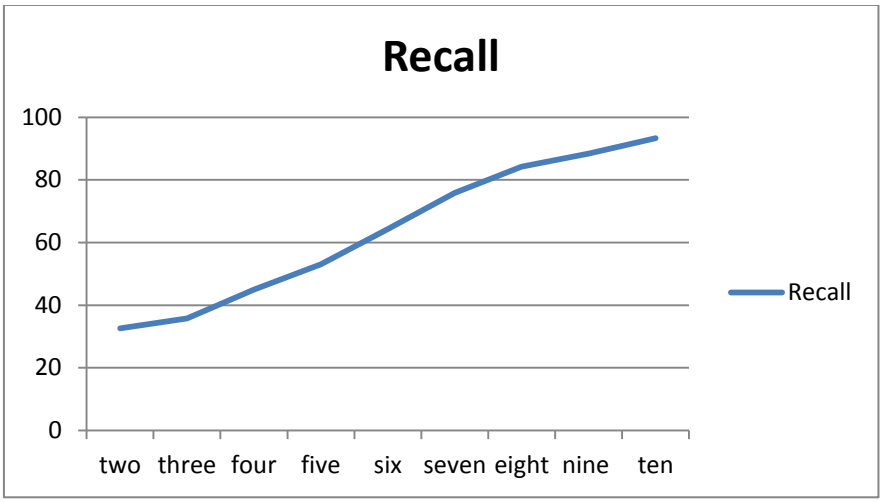

Fig 4.15 Recall of Hybrid UPGMA and K-Mean

Cluster	Recall
Two	32.6173761
Three	35.814063
Four	44.9560449
Five	53.0576527
Six	64.3057821
Seven	75.8427357
Eight	84.275227
Nine	88.3747153
Ten	93.3618502

Table 4.15 Representing Hybrid with Cosine Similarity Recall values

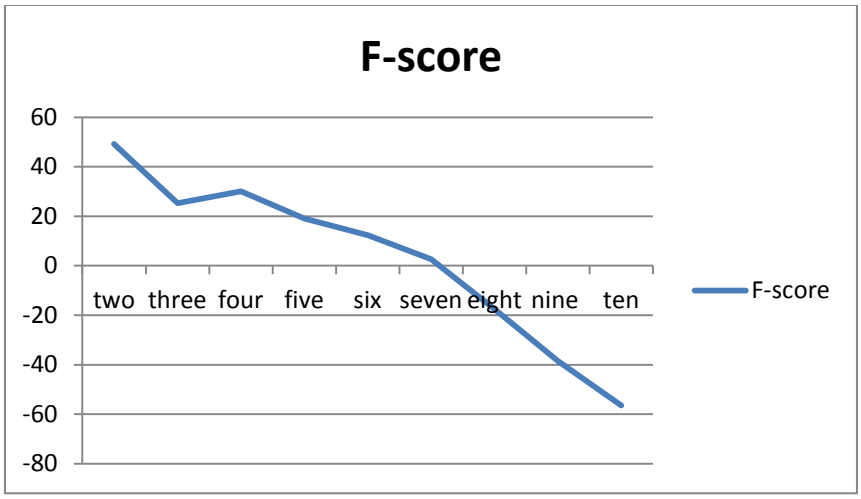

Fig 4.16 F-score of Hybrid UPGMA and K-Mean

Cluster	F-score
Two	49.1904549
Three	25.2845265
Four	30.1149433
Five	19.0944752
Six	12.3384178
Seven	2.59047609
Eight	17.5817412
Nine	38.4555023
Ten	56.4259539

Table 4.16 Representing Hybrid with Cosine Similarity F-score values

CHAPTER 5
CONCLUSION AND FUTURE WORK

In our research work we extract the texts from different sources and integrate these texts by using clustering approach. We followed four algorithms K-Mean with Euclidean distance and UPGMA with Euclidean distance and enhance version of these algorithms i.e. K-Mean with cosine similarity and Hybrid approach UPGMA and K-Mean with cosine similarity. After extracting the texts from different sources we applied the vector space model which will help in calculate document weight age and represent these document weight age in vector form. On the calculated document weight age we have applied K-Mean and Hybrid approach UPGMA and K-Mean with cosine similarity and calculate the four metrics accuracy, recall, precision and f-score.

In this experimental work we analyses the following important observations: -

1) K-Mean with Euclidean distance and UPGMA with Euclidean distance its accuracy, precision and f-score are not representing significant results but result of recall metric in both algorithms produce more effective results as compare among all metrics.
2) K-Mean with cosine similarity and Hybrid approach UPGMA and K-Mean with cosine similarity represent more better results in all the metrics i.e. accuracy, precision, recall and f-score.

Hence we conclude that outcome of K-Mean and Hybrid approach UPGMA and K-Mean with cosine similarity are much better than K-Mean and UPGMA with Euclidean distance.

In this research work we include only texts to integrate from diverse source. In the future work, we will include special characters in the texts for integration process and measure their performance based on the metrics i.e. accuracy, recall, precision and f-score.

CHAPTER 6
LIST OF REFERENCES

I. BOOKS

[1]Jiawei Han and Micheline Kamber (2013), Data Mining Concepts and Techniques.

[2] Han, Kamber and pei. (2012), Data Mining: Concept and Techniques, Morgan Kaufmann Publisher, Waltham USA.

II. RESEARCH PAPERS

[3] Lee, Mong Li, et al. "XClust: clustering XML schemas for effective integration" *Proceedings of the eleventh international conference on Information and knowledge management* ACM, 2002.

[4] Heß, Andreas, and Nicholas Kushmerick "Learning to attach semantic metadata to web services" *The Semantic Web-ISWC 2003* Springer Berlin Heidelberg, 2003 258-273.

[5] Alon Halevy, "Why your data won't mix", (2005) *Queue* 3 (8).

[6] Hakimpour, Farshad, and Andreas Geppert "Resolution of semantic heterogeneity in database schema integration using formal ontologies" *Information Technology and Management* 6.1 (2005): 97-122.

[7] Searls and David B, "Data integration: challenges for drug discovery", *Nature reviews Drug discovery* 4.1 (2005): 45-58.

[8] Hao, Guoshun and Jianghua LV, "Dynamic description logic model for data integration." *Frontiers of Computer Science in China* 2.3 (2008): 306-330.

[9] Morris, Huong, et al. "Bringing Business Objects into Extract-Transform-Load (ETL) Technology" *e-Business Engineering, 2008 ICEBE'08 IEEE International Conference on* IEEE, 2008.

[10] Shaofeng Liu, Alex H. B Duffy, Robert Ian Whitfield and Iain M. Boyle, "Integration of decision support systems to improve decision support performance." Knowledge *and Information Systems* 22.3 (2010): 261-286.

[11] Ghosh, Samiran,Saptarsi Goswami, and Amlan Chakrabarti "Outlier detection from ETL Execution trace" *Electronics Computer Technology (ICECT), 2011 3rd International Conference on* Vol. 6 IEEE, 2011.

[12] Kumar, Varun, and Nisha Rathee "Knowledge discovery from database using an integration of clustering and classification" *International Journal of Advanced Computer Science and Applications (IJACSA)* 2.3 (2011).

[13] Wang, Shaowei, David Lo, and Lingxiao Jiang, "Inferring semantically related software terms and their taxonomy by leveraging collaborative tagging" *Software Maintenance (ICSM), 2012 28th IEEE International Conference on* IEEE, 2012.

[14]Bellatreche, Ladjel and Robert Wrembel. "Special Issue on: Evolution and Versioning in Semantic Data Integration Systems", *J. Data Semantics* 2.2-3 (2013): 57-59.

[15] Ghosh, Soumi, and Sanjay Kumar Dubey "Comparative analysis of k-means and fuzzy c-means algorithms" *International Journal of Advanced Computer Science and Applications (IJACSA)* 4.4 (2013).

[16] Leida, Marcello, Alex Gusmini, and John Davies "Semantics-aware data integration for heterogeneous data sources." *Journal of Ambient Intelligence and Humanized Computing* 4.4 (2013): 471-491.

[17]Maitah, Wafa, Mamoun Al-Rababaa, and Ghasan Kannan "Improving the Effectiveness of Information Retrieval System Using Adaptive Genetic Algorithm" *International Journal of Computer Science & Information Technology* 5.5 (2013): 91-105.

[18] Gill, Rupali, and Jaiteg Singh "A Review of Contemporary Data Quality Issues in Data Warehouse ETL Environment." (2014).

III. WEBSITES

[19] http://www.ncsu.edu/labwrite/Experimental%20Design/accuracyrecallprecision.htm

[20] http://en.wikipedia.org/wiki/Scikit-learn

CHAPTER 7

APPENDIX

Glossary of Terms

A:

Algorithm: A step-by-step systematic procedure to solve some problem.

E:

Efficiency: Efficiency should be measure in term of quantitatively, in which we can taking input and measured the output.

H:

Hub: It can be defined as the point where the number of routes meets and distributed the traffic for the networking.

O:

Ontology: It is the type of constraints used in the data integration techniques. It is the WHERE clause to select the particular information from the table.

S:

Semantic: Semantic define the meaning of the data.

Abbreviations:

AGA: Adaptive Genetic Algorithm

CLIQUE: Clustering in Quest

ETL: Extract Transform and Load

DSS: Decision Support System

IDSS： Integration of Decision Support System

KDD： Knowledge Discovery in Database

NLTK: Natural Language ToolKit

NLP: Natural Language Processing

Sk learn: Scipy Toolkit

STING: Statistical Information Grid

TF-IDF: Text frequency-Inverse document frequency

TP: True Positive

TN: True negative

WEKA: Waikato Environment for Knowledge Analysis

URL: Uniform Resource Locater

UML: Unified Modified Language

UPGMA: Unweighted Pair Group Method with Arithmetic Averages

XML: Extensible Markup Language