RESEARCH ARTICLE OPEN ACCESS

Document Clustering Through Map Reducing – A Hueristic Approach

Priyanka Paradkar, Rachana Dhannawat
Department of Computer Science
UMIT, SNDT Womens's University
Mumbai

ABSTRACT
Increasing digital data drastically raises alarm for better data handling and scrutinization Technique.Numerous classification and clustering techniques exists which performs well on numerical data. But holds no good for semantic textual data, where every files or documents are correlated semantically. Certain techniques are available in industry to do document clustering based on semantics of data but most of them suffer from time complexity issue.

As an initial step to solve this, proposed methodology put forwards an idea of document clustering for different extensions like doc, pdf and txt by preprocessing the original data to perform feature extraction. To enhance the performance of process, features of documents are extracted using map reduce streams. Further these features are fed to weighted matrix method for clustering where in this process is catalyzed by fuzzy logic.

Comparative Research evaluation is been done against K-Means and K-Mediod Algorithms. Performance Graph depict our method to effective and simplified.

Keywords :- Preprocessing , Map reducin . Feature Extraction, Fuzzy logic

I. INTRODUCTION

Advancement and growth in Technology has lead to digitization vision. Number of hand written document are been currently replace by eversion at government offices and in day to day life. In digitization every piece of information is been maintained in digital format, this has greatly reduced requirement of manual intervention and paper work. As human efforts have been reduced, maintenance of information has been overhead on computer systems decreasing performance. Outcome of electronic version of articles has greatly increased electronic information and ultimately memory requirement. Although advancement in Hardware technology larger memory requirement has been fulfilled, still better software i.e. approach is been required to handle this huge information in better and intelligent way. Analysis and searching relevant information from huge heap of documents is time taking and lengthy process. To overcome this challenge document clustering is approach that has been adopted. Automated analysis of document and categorization in relevant category is been achieved with document clustering. Document clustering is a procedure of automated analysis of articles depending on type of document. Large research work exists to achieve this task. Clustering reduces time complexity with enhancement in accuracy. Document clustering approach has four layers of processing or techniques as mentioned below:

[1]Pre-Process

[2] Attribute Extraction.

[3] Master Matrix.

[4] Fuzzy Logic.

Pre-Processing is a practice of eradicating additional useless piece from information. Usually it used to reduce expenditure and time intricacy of scheme. Information pre-processing is an frequently abandoned footstep but extremely significant and is of major significance since information pre-processing outlines groundwork step of additional analysis and processing of information. Information pre-processing engrosses subsequent five layers**.** Figure 1 describes above five layers.

1. **Information cleaning**: This layer has procedure like filling in absent value, eradicating noisy information, recognizing or eradicating outliers, and solving discrepancy.
2. **Information Integration:** Layer engrosses combining information via numerous databases, data clusters, or records.
3. **Information transformation**: This layer normalization and aggregation process is performed on information that has been combined from diverse sources.
4. **Information reduction**: This Layer size and volume of information is be compressed to similar logical outcome.
5. **Information discretization**: This layer Information of pre-processing discretization process like restoring arithmetical features with insignificant ones.

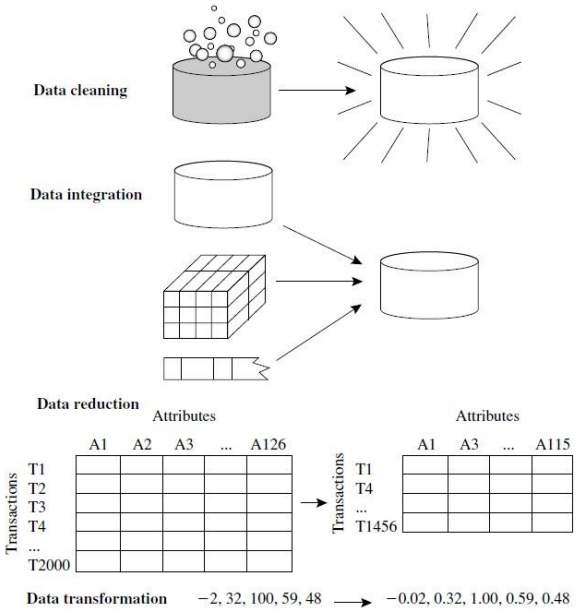


Fig 1: layers of Pre-Procesing[1]

Feature Mining

As identified in above section data mining is a practice of information extraction from huge deposit of Information with assistance of certain programmed technique that would reduce expenditure and time process complexity of system. So assumed procedure has relevance in array of region that as AI(artificial intelligence), ML(machine learning), HCI(human computer interaction) .classically data mining assignment address to discover a unidentified value of feature from its recognized values.

Numerous of data mining Information is present in multi-dimensional outline and could be best characterizing using numeral of attributes. Additional dimensions Information has more would be cost complexity of information. As such usually dimensionality lessening procedures are castoff before information is been given for classification.

Feature mining is one of better technique for dimensionality lessening and is been under practice. In feature mining method a new subgroup of attributes are mined from unique attribute set by maintaining as much of data as in unique data as probable. Numerous true world applications are there that have numerous attributes and these attributes can be used for classification. But if all these attributes are worn then perceptibly it would give rise to huge expenditure and relevance generation complexity. Once dimensions are concentrated toughness of classifier would be developed and cost difficulty would be condensed.

Feature assortment is a method to find first-class superiority of pertinent features from unique mining dataset using certain information reduction and attribute mining procedures. Feature mining engrosses assortment of a attribute termed as Feature assortment, Feature choice step has turn out to be a demanding subject in field of Pattern identification , Machine and artificial intelligence, Data Mining and Case Based logic .Feature assortment is a procedure of discovery of an best or sub-best division of 'N attributes from innovative 'N' attributes. It necessitates an outsized search space to acquire attribute division. Optimal attribute separation is considered by assessment criteria. Chief intention of attribute choice is to decrease numeral of attributes and to eliminate irrelevant, unnecessary and noisy Information.

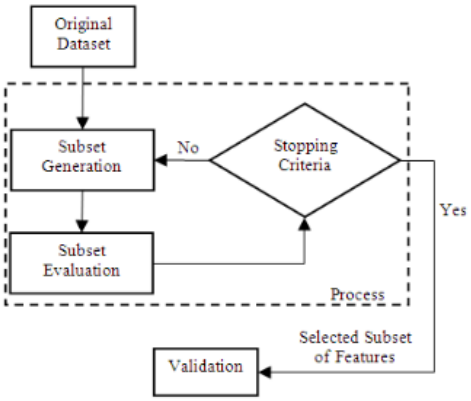


Fig 2: Feature extraction steps[2]

Figure 2: explains various steps involved in attribute mining procedure.

Fuzzy Logic

Most widely and largely used research algorithm is for classification of information is fuzzy logic. It evolves four layers of processing.

I. Fuzzification unit:

Conversion of system input into crisp facts that is fuzzy sets is processed here. To achieve this task relationship function accumulated in fuzzy data base is used in here and fuzzy linguistic variable is worn for superiority presentation.

Straddling a particular range or cross area. It too splits and put sign in as five steps:

LP	**x is Large Positive**
MP	**x is Medium Positive**
S	**x is Small**
MN	**x is Medium Negative**
LN	**x is Large Negative**

II. Fuzzy Knowledge Base unit:

Rule based condition set by user is set in logic of IF-ELSE. It is totally pedestal on association and linguistic Methods.

A.] Linguistic Variables: variables signifying in natural languages as a substitute of numeric value.

B.] Association Functions: major task of Association method is to map non-fuzzy input to Fuzzy input.

III. Inference Engine unit:
This Layer Human logical reasoning intelligence step is replicated by scheme by Generating fuzzy inference engine with IF-ELSE logic.

IV. De-ffuzzification unit:
Previous layer generated fuzzy set is used here to form crisp values.

As stated data mining is process of mining frequent items from dataset. As huge volume of information is present larger time is required. Map Reduce approach compresses time acquired to accomplish task as numerous tasks are performed simultaneously.

The successive sections of this paper are dedicated as section 2 for Literature survey, Section 3 for proposed methodology of our system. And detailed results and comparisons are discussed in the section 4 and Section 5 for Conclusion and Future scope this paper.

II .LITERATURE SURVEY

A. Survey Technique

Analyzing and reviewing every technique under Document clustering process is technique adopted for survey.

B. Survey

Method of performing preprocessing Along with association rules have been explained by[2]. Spatial rules of association have been found with apriori and FP-growth. Future this research work can be expanded to effective preprocessing. Research evaluation of proposed preprocessing algorithm present that it is time consuming and major reason is spatial association rule process.

Feature mining can also be termed as dimension reduction technique. Article presented by [3] focuses on Common research issue faced by this mining algorithms that is dimensionality although there is no perfect classification procedure it's most challenging task to design full fledge automated classification system for mining attributes.

Three foremost vital techniques of attribute extraction for classification have been implemented based on Eigen vector in research work [4]. Proposed techniques have been tested for three applications. Numerous bottlenecks have be found frequent discontinuity in generated results. So to conquer limitation author has made numerous changes and assumptions.

Automated technique of attribute extraction hereditary from current text classification technique to map bigger dataset to manageable dataset is developed in [1] . Figure 3 below defines text categorization scheme. Enhanced version of scheme represented as above in figure 3 is been used for commercial purpose.

Image handling library that could be used more efficiently along with framework of MAP reduce i.e HIPI is been described by research work[5]. Complex computations have been considered specially in designing system. Abstraction layer has been maintained and focus remains on image processing techniques.

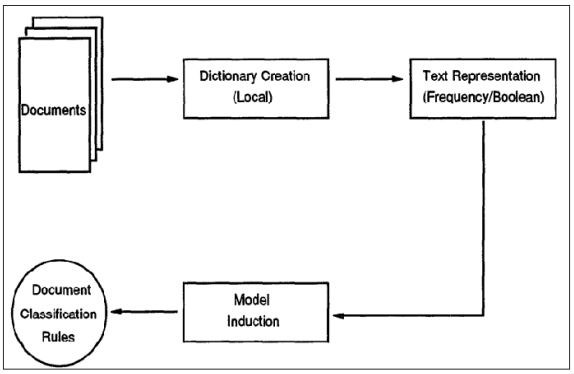


Fig 3: Text categorization model [4]

In case of user image is always major concern hence abstraction layer is implemented. Observations reveal as information increases throughput of system increases i.e with increase in image set output is better. MAP-Reduce provides different provisions to process image. In MAP-Reduce culling process filter out different images. Decoder and Encoders mechanism is implemented in background to generate float images.

In generalized view complete view of framework on integration is as shown in figure 4 and is easy to mechanize.

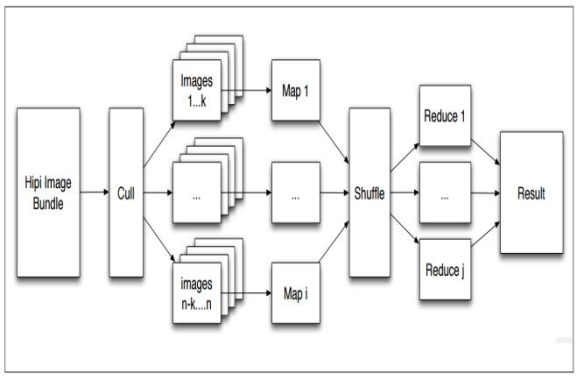


Fig 4: pipeline architecture of Map-Reduce Framework [6]

Fuzzy Mechanism for E-learning system has been proposed in [6].Scheme is been used to learn a language, with various phases like memory cycle updates, Fuzzy mechanics, AHP etc. This technique helps in analysis of articles to best understanding ability of consumer.

As a consumer reads article a test is been conducted which assist in increasing vocabulary of words. This newly keyword enhances dictionary of scheme and make easy to learn in e-learning system. Above method consumer dislikes

and likes are considered and in accordance to consumer levels of preference best articles are been given to them. Dictionary is been recalculated for words after every process.

Word document frequency with matrix technique is complex and generates huge space with sparse vector for mining is been observed by[7]. Proposed work is based on vector of Non-negative factorization matrix. Result evaluation show that features are compressed from 1000 to 100 set and successful document cluster generation.

Hybrid approach on clustering document from web information is been proposed in [8]. Author suggests classification of information is required in order to reduce search time .proposed system uses concept based procedure with AHC to organize cluster. Generated cluster are based on both content analysis and urls examination.

Text clustering approach based on association rules based for hierarchical clustering is been tested on Reuters dataset of 2157 and shows enhanced cluster quality in research work of [9].

Two issues related to document clustering, time complexity and lower relatedness affects clustering efficiency. Proposed approach of [10] is hybrid which combines attributes from DRR(Directed Ridge Regression),Fuzzy clustering and conceptual clustering. Evaluated work show that clustering results are better and accurate in comparison to all of existing techniques.

III PROPOSED METHODOLOGY

The proposed methodology of document clustering technique can be broadly explained in this section with respect to the mentioned steps of the figure **5** with the following steps.

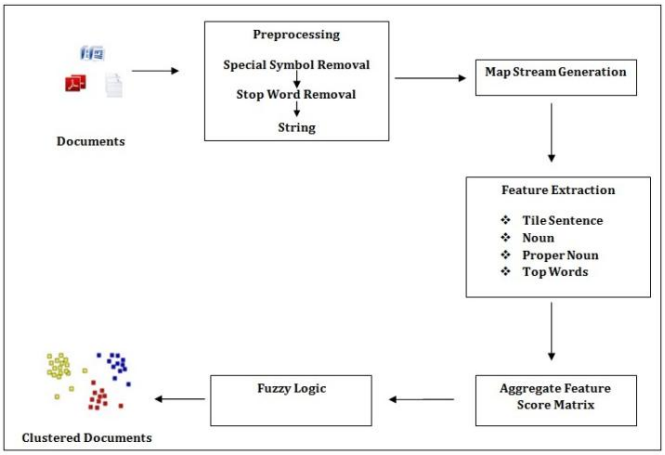


Fig 5 : System Overview

Step 1: Here in this step a folder is given as input for document clustering process which contains different files of different extensions. But the proposed system is built to work on textual files with the extensions like .txt, .Doc and .pdf. Once the System is feed with these files as input it read all the file content as string and store in a list.

Step 2: Preprocessing of the string eventually lessens the burden on the process by maintaining the quality of the data. Due to this process all the redundant data will be eliminated so that the whole process of document clustering will be run with lesser time and space complexity. This process involves several steps as narrated below.

- **Special Symbol Removal**: This process eventually drops all the special symbols from the string of the files. Symbols are like,,.,:,;,?,{,},[,] etc…

- **Stopword Removal:** This is the process of discarding the conjunctions from the file string. By achieving this, the meaning of the string never changes. For example if a String is like "we celebrated Diwali", After the Stopword removal process string becomes "celebrated Diwali". In this process the semantic of the sentence is retained even though it shred off the conjunction words. For this system stores the static information of all the available stopwords in the English language and then by linear comparison of each words it shreds off the each conjunction words.

Stemming: Stemming is the process of bringing the word to its base form by trimming the postfix tenses. For more understanding of this let the stopword removed word from the prior step will consider, i.e. "Celebrated Diwali". After Stemming of each word String becomes" Celebrate Diwali". Even by doing this also sentence is capable of retaining its original meaning.

To achieve stemming many algorithms are existed like Port stemmer, Stans Stemmer and many more. Each of these algorithms needs their own protocol to trim the word and it may not suitable for all kinds of textual data. And some universal stemming tools are available like Google stemmer, which is indeed a costlier affair for any protocol development.

So the proposed system uses the string replacement technique for trimming of the unwanted tenses from the words. Here predefined protocols for the tenses are used for efficient trimming process.

By doing all this preprocessing process certainly yields the light weight textual data for the further process of document clustering.

Step 3: This step of map reduction eventually boosts the process of document clustering by creating the multiple partitions of the documents. Then each of the partition of the documents is feed to the multiple threads in run time for parallel computation of the feature extraction process from the documents, which leads to the semantic clustering in the future steps. This process can be bitterly seen with the below mentioned algorithm.

ALGORITHM1: Document Partition for MAP Streams

// Input: Documents List
// Output: Partitioned Lists
Step 0: Start
Step 1: Decide partition number N

Step 2: Declare an Empty List L
Step 3: Read all the documents names in a vector D
Step 4: Number of Map Stream M= Size of D/N
Step 5: Set Count=0
Step 6: For i=0 to size of D
Step 7: count++
Step 8: Add D_i into Temp List T
Step 9: IF count=M
Step 10: Add T to L
Step 11: Reset T
Step 12: END FOR
Step 13: return L
Step 14: Stop

Step 4: Extraction of features from the text finally yields the more meaningful facts of the documents in desired manner than the unstructured data. For the clustering of documents in highly semantic manner, feature of the textual data need to be separated in proper manner with the following aspects

- ✓ Title Feature **(T_F)**
- ✓ Numerical Data Feature **(N_D)**
- ✓ Proper Noun Feature **(P_N)**
- ✓ Top Word Feature **(T_W)**

- **Title Feature:** As the name itself is narrating about this feature it includes the very first sentence of the documents as the title sentence. System collects all these words and store in a List.

Title of any documents is playing the crucial role in understanding it in very meaningful way. And consideration of this eventually adds a feather in the semantic clustering of the documents.

- **Numerical Data:** Numerical Data of any document is as important as the whole documents itself, as numeric are playing very vital role in deciding documents quality. So Consideration of the numerical data feature for document clustering is good idea of enhancing the quality of the clustering.

Proposed system takes all the preprocessed string of the documents and places all words in a vector. Then System considers every word and tries to convert into a numeric data type. If any value is not able to convert into numeric data type (Integer or Float) and it yields any exception that means the word is pure String else it is a numeric data. If it is numeric data then it collects it into a vector for further processing.

- **Proper Noun:** Proper nouns in any textual data immediately catches attention, it is quite obvious that it always represents a living entity or place. So extraction of this as feature for documents clustering directly affects the quality of the clustering in all sense.

For this purpose proposed system uses a dictionary of words in MS- Excel format which is collected over the internet. This dictionary contains more than 100 thousand words of English language except the Proper nouns.

So to identify the proper noun system puts all the pre-processed words in a vector and then every word is been checking for its presence in the dictionary. If the word is not present in the dictionary then the word is considered as the proper noun else it is not.

- **Term Weight:** In simpler terms it is also referred as the top words, they are actually representing the most repeated words in the text. These top word are indicating the importance of the words that in turn makes the whole document itself important.

Proposed system firstly identifies all the words from the documents for their uniqueness and adds all to a vector. Then for every word from the vector system calculates the frequency for each of them. And then by sorting in descending order till user defined size system gets top words for the respective documents.

Step 5: Once all the features are extracted for the input documents by the map streams of parallel computing, then all these features are bonded with the respective documents in a list. This list is arranged in rows and columns to yield the mean aggregate values of the features scores by comparing with all other remaining documents. This mean aggregate value indeed represents a weighted matrix of the features which finally help the system to get highly semantic clusters of the input document. This whole process can be clearly observed in the below mentioned algorithm.

ALGORITHM2: Aggregate Feature Score matrix Generation For document Clustering

// Input: Documents Feature matrix D_F
// Output: Clustered Documents
Step 0: Start
Step 1: Initiate a Feature matrix M_F of size Number of documents
Step 2: **FOR** each D_F elements
Step 3: Extract all Features
Step 4: **FOR** each D_F elements
Step 5: **IF** Index are not equal
Step 6: Set Matching scores M_s=0
Step 7: **ELSE**
Step 8: Get all Features for D_F
Step 9: Calculate the matching score for all features
Step 10: Get the mean of the matching Score M_s
Step 11: add M_s to position of M_F matrix
Step 12: **END Inner FOR**
Step 13: **END Outer FOR**
Step 14: Stop

Step 6: Fuzzy Logic is an abstract classifier that eventually creates the classification segments known as the fuzzy crisp

values. For the created feature score matrix M_F crisp values are ranges just as follows.

- VERY LOW – 0 to 0.2
- LOW - 0.21 to 0.4
- MEDIUM - 0.41 to 0.6
- HIGH - 0.61 to 0.8
- VERY HIGH -0.81 to 1.0 .

So for each of the documents D_{Fi} in the row of the matrix MF a cluster is been formed based on the user define accuracy that are normalized to fuzzy scores with respect to the documents scores align in each columns as D_{Fj}. This process of the proposed system finally yields the fine grained semantic clusters for the given input documents.

IV RESULTS AND DISCUSSIONS

To evaluate the results and discussions of the proposed system some experiments are conducted on the windows based java machines. The performance is evaluated on Core i5 processor with 4GB of RAM by feeding documents as input to the system belongs to different text formats like pdf, txt and doc/docx files to the system.

System is put under hammer for different test as stated below.

4.1 Performance evaluation of the system:

Here in this measurement analysis process system is put under tests for its performance of cluster evaluation based on time. The results obtained for different volume of data is been recorded and depicted in the below table 1.

Document numbers	Time in seconds
5	30
10	34
15	41
20	47

Table 1: Time required against numbers of documents

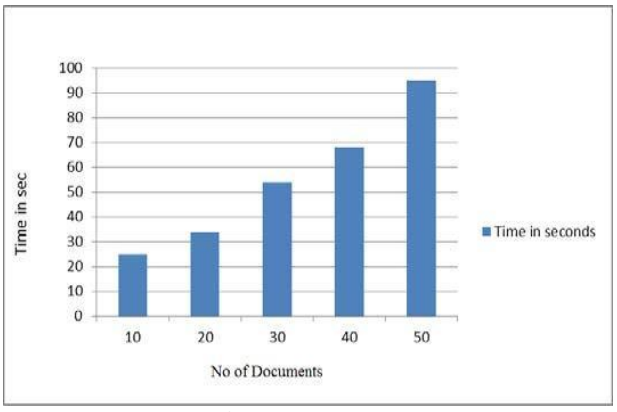


Fig 6: Performance measurement

In the above graph mentioned in the figure 6 clearly denotes that the time required for the clustering the documents are not directly proportional to the number of input documents. This indicates that system over performs for the given number of documents for clustering them.

No of Clusters	Time Taken To Execute (In milliseconds) K-Mean Algorithms	Time Taken To Execute (In milliseconds) K-Mediod Algorithms	Time Taken To Execute (In milliseconds) Weighted matrix method Algorithm
2	21345	20328	14044
3	42175	40385	21067
4	68512	69214	28089
5	74365	73343	35111

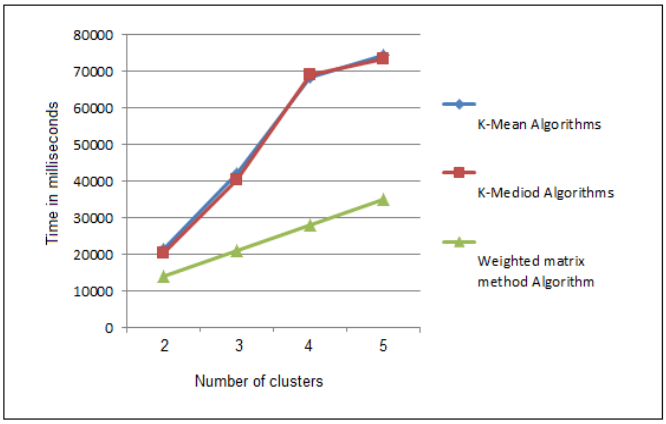
4.2 Comparison with K-Means and K- mediod

When our system is compared for the performance time for the number of clusters extracted for the given data with k means and k-mediod algorithms with our weighted method matrix algorithm we found some facts , which are depicted in the below table 2.

Table 2: Time comparison table

Fig 7: Comparison Graph with K-means and K mediod algorithms

On drawing the plot for the table 2 we got the resulted



graph as mentioned in the figure 7. The plot clearly indicates that our method of weighted method matrix technique over performs than the k means and k mediod clustering that mentioned in [x].

4.3 Comparison Single Link – Complete Link

When our system is compared for the performance time for creation of clusters for the given data with single link and complete link algorithms with our weighted method matrix algorithm we found some facts , which are depicted in the below table 3.

Table 3: Time Comparison table

Sr No	Algorithm Techniques	Time Taken for Clustering (In milliseconds)
1	Single Link	830
2	Complete Link	570
3	Weighted matrix method	498

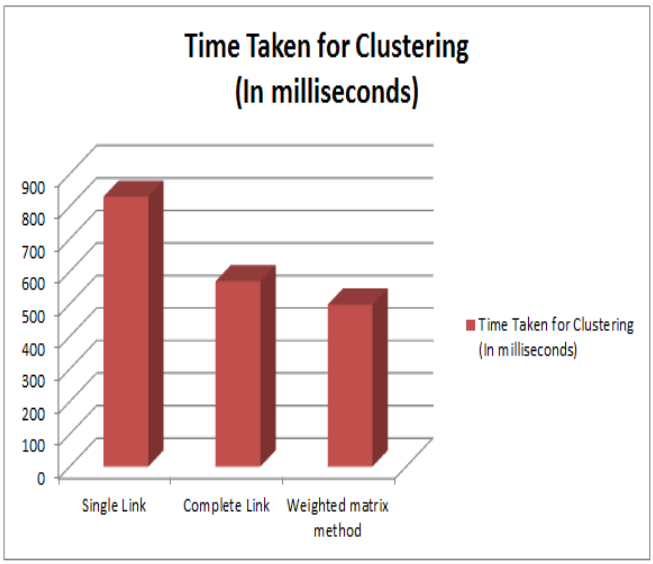


Fig 8: Comparison Graph with Single and Complete Link methods

On drawing the plot for the table 3 we got the resulted graph as mentioned in the figure 8. The plot clearly indicates that our method of weighted method matrix technique over performs than the other two in performance time of clustering as stated in [y].

V. CONCLUSTION AND FUTURE SCOPE

The proposed system successfully incorporates the process of feature extractions using Map reducing technique. System efficiently fetches the feature from the documents like

- ✓ Title Sentence
- ✓ Numerical Data
- ✓ Proper Noun
- ✓ Top Words

And the system efficiently scrutinizes the semantic of the data by weighted method matrix process which yields the fine clusters of the documents. To classify the clustered documents in more meaningful way an abstract classifier like fuzzy logic is used to get the accurate clusters based on the semantics.

This concludes in the end that system is contributed its bit to the clustering of the documents more semantically than the random clustering of the documents based on some numeric.

As future vision of this version of the system it can be enhanced to perform in the distributed paradigm for more reliable number of the documents of any sizes.

REFERENCES

[1] Jiawei Han Micheline Kamber Jian Pei, “Data Mining: Concepts and Techniques” Elsevier 3rd Edition, 22nd June 2011,744.

[2] Aniket Pote, Swapnil Sutar, Priyanka Sonawane, Fatima Inamdar, “Social Recommendation SystemTechnique:Survey” , International Journal of Emerging Research in Management &Technology ISSN: 2278-9359 (Volume-5, Issue-3), March 2016

[3] M Pechenizkiy, S Puuronen, A Tsymbal, “Feature Extraction for Classification in the Data Mining Process”, nternational Journal "Information Theories & Applications" Vol.10,2007

[4] Jasdeep Singh Malik, Prachi Goyal, Mr.Akhilesh K Sharma, “A Comprehensive Approach TowardsData Preprocessing Techniques & Association Rules” Bharati Vidyapeeth's Institute of Computer Applications and Management (BVICAM)/news/Indicompapers, 2009.

[5] Chris Sweeney, Liu Liu, Sean Arietta, Jason Lawrence, “HIPI: A Hadoop Image Processing Interface for Image-based MapReduce Tasks”, Research gate BS Thesis. University of Virginia, Department of Computer Science 2011

[6] Tung-Cheng Hsieh, Tzone-I Wang Chien-Yuan Su and Ming-Che Lee , “A Fuzzy Logic-based Personalized Learning System for Supporting Adaptive English Learning” , International Foru m of Educational Technology & Society (IFETS). ISSN 1436-4522 (online),2012.

[7] paresh chandra barman,Md. Sipon Miah, Bikash Chndra Singth,“Feature extraction clustering in text mining using NMF basis probability”, Ulab journal of science and engineering ,november 2,2011.

[8] Gamare, Pralhad S., and G. A. Patil. "Web Document Clustering using Hybrid Approach in Data Mining." International Journal of Advent Technology 3.7 (2015).

[9] Rose, J. Dafni. "AN EFFICIENT ASSOCIATION RULE BASED HIERARCHICAL ALGORITHM FOR TEXT CLUSTERING." International Journal of Advanced Engineering Technology Vol. VII/Issue I/Jan.-March 751 (2016): 753.

[10] Nagaraj, R., and X. Agnise Kalarani. "Semantically Document Clustering Using Contextual Similarities." International Journal of Applied Engineering Research 11.1 (2016): 71-76.