

# Establishing Hidden Semantics in Web Documents Using Fuzzy Clustering By Feature Matrix Approach

<sup>1</sup>Aditya Deshpande, <sup>2</sup>Dr. Pramod Patil

<sup>1</sup>Research Scholar, Dept. of Computer Engineering, Pad.Dr.D.Y.Patil Institute of Engineering and Technology, Pune, Maharashtra, India

<sup>2</sup>Professor & HOD, Dept. of Computer Engineering, Pad.Dr.D.Y.Patil Institute of Engineering and Technology, Pune, Maharashtra, India

## Abstract

As the data is exponentially exploding on the 'World Wide Web', the conventional clustering algorithms obligate various challenges to tackle, of which the most often faced challenge is the uncertainty. Web documents have become heterogeneous and very complex. There exist multiple relations between one web document and others in the form of embedded links. This can be imagined as a one to many (1-M) relationship, for example a particular web document may fit in many cross domains viz. politics, sports, utilities, technology, music, weather forecasting, linked to e-commerce products etc. Therefore there exists a necessity for efficient, effective and constructive context driven clustering methods. Conventional clustering algorithms adhere to classify the given data sets as exclusive clusters. Signifies that we can clearly state whether to which cluster an object belongs to. But such a partition is not sufficient for representing in the real time. So, a fuzzy clustering method is presented to build clusters with indeterminate limits and allows that one object belongs to overlying clusters with some membership degree. In supplementary words, the crux of fuzzy clustering is to contemplate not only the belonging status to the clusters, but also to cogitate to what degree does the object belong to the cluster.

**Keyword:** Clustering, Fuzzy Logic, Weighted Matrix, Fuzzy Clustering, Conventional Algorithms, Domains, Data Driven

## Introduction:

With an incredible circulation of several hundred million sites worldwide, the ever changing cluster of web documents is getting bigger and bigger every day. This incorporates some very important and as well very difficult challenges. Over the preceding duration of ten years there has been incredible growth of data on World Wide Web. It has become a major source of information. Internet web generates the new defies of information retrieval as the amount of data on the web and number of users using web growing rapidly. It is challenging to quest through this tremendously large catalogue for the information desired by user. Henceforward the need for Search Engine ascends. Search Engines use crawlers to collect data and then store it in database maintained at search engine side. For a given user's query the search engines searches in the local database and very quickly displays the results.

## Web Crawlers:

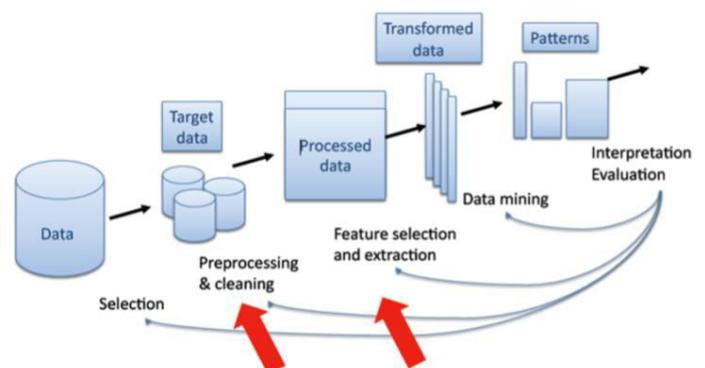
Web crawling is an imperative method for amassing data on, and custody up with, the speedily intensifying Internet. Web

crawling can likewise be baptised as a graph search problem as web is considered to be a large graph where nodes are the pages and edges are the hyperlinks. Web crawlers can be used in various areas, the most prominent one is to index a large set of pages and permit other people to search this index. A Web crawler does not really move all over the place on the computers linked to the Internet, as viruses or bot agents do, as a substitute it only directs treaties for documents on web servers from a set of already sites.

However the web crawlers have progressed, there has remained significant weakness in search engines outstanding to the complex, inter related (linked or cross domain documents) in the document assembly. Polysemies, synonyms, homonyms, phrases, dependencies and spam's act as hindrance to the search engines and therefore hampering the results returned. Also the vagueness or irrelevance of the user probes increases the ambiguous results fetched.

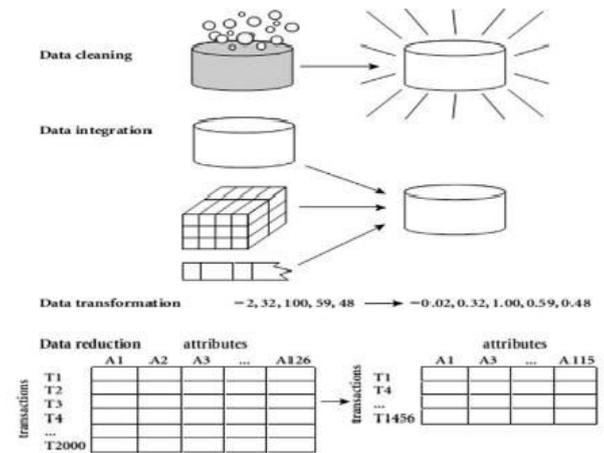
**Pre-processing:** Data pre-processing exists an often neglected step but very important and is of prime importance since data pre-processing forms the foundation step of further analysis and processing of data. Data pre-processing involves following five steps:

1. Data cleaning: This step has operations like to fill missing values, smoothing out the noisy data, detecting or eliminating outliers, and deciding discrepancies.
2. Data integration: It involves integrating data using numerous databases, data cubes, or collections.
3. Data transformation: In this step we perform normalization and aggregation operations on data which has been integrated from various data sources.
4. Data reduction: In this step we condense the volume of data and produce the similar investigative results.
5. Data discretization: In this step of data pre-processing we perform discretization operations like replacing numerical attributes with nominal ones.



The phrase – If you input the junk data that is ‘Garbage In’, then you be surely getting the junk output that refers to ‘Garbage Out’ is particular to the domain of machine learning. Data congregation methods are often range values, irregular, missing values. Analysing data that hasn’t been properly processed, such data can produce misleading results. Thus, pre-processing is primarily important step formerly running an investigation. Data fetched by a web crawler needs significant amount of processing before it is fed to the ‘Fuzzy Clustering Algorithm’ (FCA). Data in actual world is unclean which means it is incomplete. Incomplete data means it lacks attribute or the data in which we are interested in. The second part of dirty data is that it contains noise. Noisy data means that there are inaccuracies or outliers in it. The third part of dirty data is that it is inconsistent. Inconsistent means that the facts is not in correct format or the data lacks proper coding and naming format. If there is no good quality of data available then the data that would be eventually loaded in the data warehouse would be of low standards. Meaning that the mining algorithms would yield a junk result out of the data warehouse. For data to be in correct format for data mining it should possess some valuable qualities wherein the data mined would be of highest quality. These desirable qualities are precision, reliability, comprehensiveness, attribute value and most importantly timeliness. The most vital part of Pre-processing is cleaning the data. If the right data is not fed in we cannot expect the right output. Therefore cleansing of data is most vital. Missing data means adding the missing values. Adding missing values means filling the missing values with the average value derived by mean method. It also has a step of removing the noisy data. As discussed above noisy data means the data which contains errors or outliers. Data cleaning also involves removing of inconsistencies. Inconsistent data removal means removing the data which falls into outlier range. Data can be collected from multiple formats like different databases, different file formats. The most important part is to collate this data and then cleansing this data. It involves converting the dates to one particular format, converting numeric data to proper decimal format. It also involves performing binning on the numeric data. Filling the missing data is done by usually adding a tuple or replacing the missing data by a global constant. Applying the data cleansing task in my work the first step is to remove the stop words from the data which has been fetched by the web crawler. Removal of stop words and stemming: In this phase, data which has less semantic is removed. Meaning, the full stops, commas, conjunctions etc. are removed. The data of the web documents fetched by the web crawler is compared against a bag of words – Stop words. The matched records are eliminated from the data file. Stemming is a pre-processing step making the data ready for the next step. In fact it is very important in most of the Information Retrieval systems. The main perseverance of stemming is to decrease different grammatical forms or the word forms of a word like its noun, adjective, verb, adverb etc. to its root form.

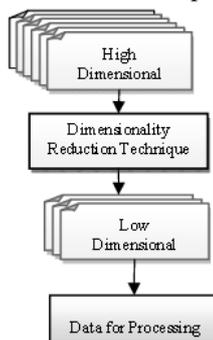
The goal of stemming is to diminish devotional forms and occasionally derived forms of a word to a conjoint base form. The available data is now further processed.



The next step of data pre-processing is integrating data from various databases, files, cubes together. Data integration means combining data into intelligible store. It also means schema integration. Schema integration means integrating metadata from different sources. It also involves identifying the attribute mismatch and performing a correcting action. This data is then fed to Data Reduction engine. In data reduction, the data which is redundant is removed. Data redundancy occurs because data is integrated from multiple databases, files etc. It means that same attribute might have been referred in a different way or the value of same attribute would have been derived or calculated. Redundant data is recognized by co relational analysis. Data reduction also involves performing numerosity reduction on data. It means to apply the linear regression model on data. This step is followed by Data Transformation. Data transformation means transforming the data in a format which is consistent throughout. It involves normalizing the data and aggregating it. The smoothening process of data transformation removes the noise from data. Aggregation step means aggregating the data into summarized cubes. Normalization activity means scaling the data such that it falls under particular range. It also involves construction of new attributes. It states that the data now has been fully transformed and ready to be loaded in the warehouse. We can conclude that data preparation is a critical issue for both data warehousing and data mining, as actual world data tends to be imperfect, noisy, and unpredictable. Data preparation involves data cleaning, data integration, data transformation, and data reduction. Data cleaning mechanism could be used to fill in missing values, lessen noisy data, detect outliers, and correct data inconsistency. Data integration loads data from multiples sources to form an intelligible data store. Metadata, correlation analysis, data skirmish detection, and the determination of semantic heterogeneity contribute to smooth data incorporation. Data alteration techniques conform the data into appropriate forms for mining. Data reduction methods such as dimension reduction, data cube aggregation, numerosity reduction, data compression and discretization could be used to obtain a reduced representation of the data, while minimizing the loss of information content. Concept hierarchies establish the values of attributes or dimensions into measured levels of abstraction. They are forms of discretization that is predominantly useful in multilevel mining. Automatic cohort of concept hierarchies for categoric data may be grounded on the number of distinct values of the attributes defining the hierarchy. For numeric data, practices such as

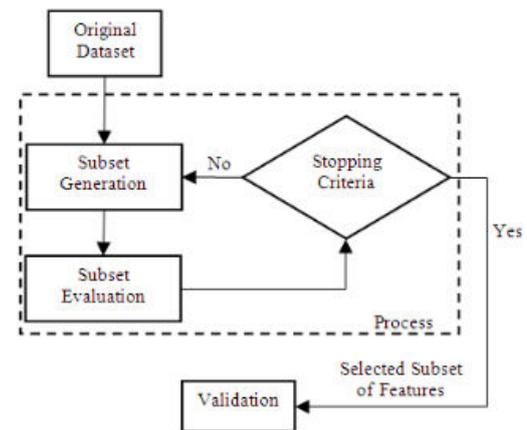
data segmentation bydividerdocumentations, histogram analysis, and clustering analysis can be used.

**Feature extraction:** Data mining is the cumulative task of data analysis and detection algorithms to perform automatic extraction of information from vast amounts of data. This process bonds manypractical areas, counting databases, human-computer interaction, statistical analysis, and machine learning.A typical data-mining chore is to forecast an unidentified value of circa attribute of a new occurrence when the valuesof the supplementaryqualities of the new occurrence are recognised and a collection of instances with known values of all theattributes is given. Most importantly in numerous applications, data is the subject of analysis and processing in data mining,is multidimensional, and presented by a number of topographies. There are moreover many dimensionsof data that it isrelevant toseveral learning algorithms, denotes the extreme raise of computational complexity and classification error with data having high expanse of dimensions. Hence, the dimensionality of the feature space is habituallyabridgedeforecataloguing is commenced. Feature extraction is one of the dimensionality lessening techniques. Feature extracts a subset of novel features from the unique feature set by means of some functional mapping possessing as much information in the data as possible. Many of the definite world applications has numerous features those are used in an effort to safeguard accurate cataloguing. If all those features are used to build up classifiers, then they function in high dimensions, and the learning process becomes complex, which leads to high cataloguing error. Hence, there is a necessity to condense the dimensionality of the features of data before classification. The key objective of dimensionality reduction is to convert the high dimensional data samples into the low dimensional data space such that the core information contained in the data is preserved. Once the dimensionality isreduced, it aidsus to improve the robustness of the classifier and also to decrease the computational complexity.



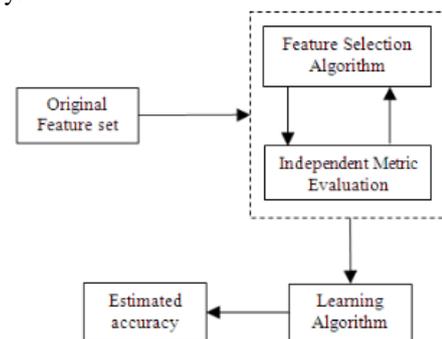
Feature assortment is a technique tofind good quality of germane features from the unique dataset using some data reduction and feature extractionmeasures. Feature extraction involves selection a feature, this is called as Feature Selection, Feature Selection step has turned out to bea thought-provokingconcern in the field of Pattern Recognition , Data Mining ,Machine Learning and Case Based Reasoning .Feature Selection is process of finding an ideal or suboptimal subset of ‘n’ features from the unique ‘Features. It requires a large search space to obtain the feature subset. The ideal feature subset is measured by evaluation criteria. The key objective of the feature selection is to decrease the number of features and to remove the irrelevant, redundant and noisy data.

Feature Selection encompasses various steps. These steps are depicted in a diagrammatic state as below

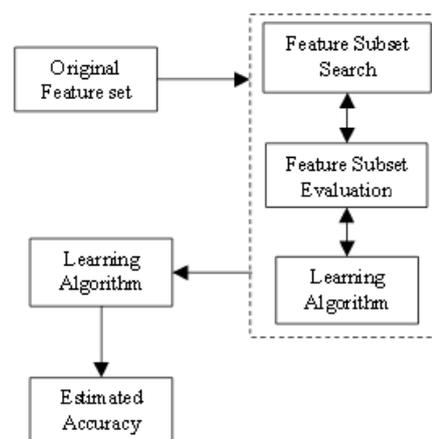


Feature selection mechanism aremostly classified into three types. They are, Filter Approach, Wrapper Approach and Hybrid Approach.

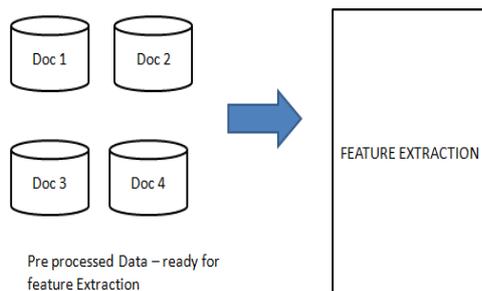
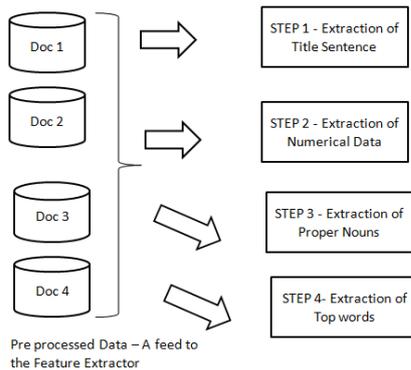
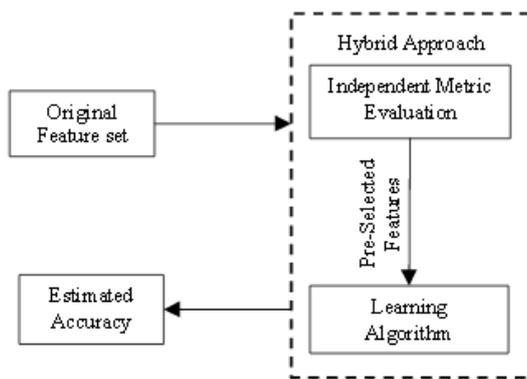
Feature selection method of ‘Filtering’s anarithmetical measure used as a criterion for selecting the relevant features. This approach is calculated easily and very efficiently.



The Wrapper methodology follows the learning algorithm is to select the relevant features from the massive dataset.



The Hybrid methodology is developed by combining the above filter methodology and wrapper methodology to handle larger datasets. In this methodology the feature set is evaluated using together independent measure and a data miningalgorithm. The sovereign measure is used to select the best subset or a specified cardinality and the data mining algorithm selects the finest subset among the best subsets across diverse cardinalities.



With the processed data now available, we now extract the important 'features' available. The first step in feature extraction process is to extract the 'Title Sentence'. The first line of the document is rendered as the 'Title Sentence'. The second step is extracting the numerical data in the data file. A scan is performed and all the numerical data present in the data files is counted. Next, all the nouns present in all the data files are listed. Only proper nouns are to be used. The last scan is done for the 'top words'. Each document now is scanned. The word whose count is highest is regarded as top word. A list of such words is made in descendant order rendering to the web documents.

**Fuzzy Logic:** Fuzzy Logic System are those which produce satisfactory but definite output in rejoinder to imperfect, vague, partial, or imprecise (fuzzy) input. Fuzzy Logic (FL) is a technique of reasoning that is similar or resembles anthropological reasoning. The methodology of Fuzzy Logic tries to inherit the way of conclusion making in humans that encompasses all transitional possibilities between digital values Yes and No. The predictable logic block that a computer can comprehend takes precise input and produces a definite output as TRUE or FALSE, which is corresponding to human's YES or NO. The creator of fuzzy

logic term, LotfiZadeh, detected that dissimilar computers, the human conclusion making embraces a range of likelihoods between YES and NO, such as : CERTAINLY YES, POSSIBLY YES, CANNOT SAY, POSSIBLY NO, CERTAINLY NO.

Fuzzy logic system comprises of four key parts: A Fuzzifier, Rule Base mapper, An Inference Engine and Defuzzifier.

**Fuzzy Logic Systems Architecture is as follows:**

**1. Fuzzification module:** This module converts the input to the systems, which are in the form of crisp numbers, into fuzzy sets. For example it transmutes the supplied crisp values to a linguistic variable by making use of the membership functions warehoused in the fuzzy knowledge base. Fuzzy linguistic variables are used to epitomise qualities straddling a particular spectrum or cross domain. It also ruptures the input sign into five stages such as:

L.P	Large Positive
M.P	Medium Positive
S	Small
M.N	Medium Negative
L.N	Large Negative

**2. Fuzzy Knowledge Base module:** It stocks the conditions established on the If and then rules provided by experts. The fuzzy knowledge base is constructed on linguistic and membership functions.

a) Linguistic Variables: Linguistic variables act as input or output for the system. Their values are articulated in a natural language as an alternate to numerical values. A linguistic variable exists as a generally disintegrated into a group of linguistic terms.

(b) Membership Functions: A membership function is used for 'quantifying' the linguistic term. Membership functions are used in the fuzzification and defuzzification steps to plot the non-fuzzy input variables to fuzzy linguistic terms and vice versa.

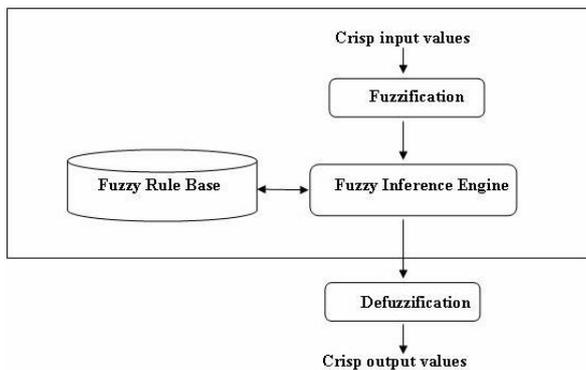
**3. Inference Engine module:** It feigns the human cerebral method by creating fuzzy interpretation on the inputs and IF-THEN rules.

**4. Defuzzification Module:** It transmutes the fuzzy set gained by the inference engine into a crisp value.

The exponential growth of the Web has led to extensive expansion of web content. The huge expanse of product information on the Web poses an inordinate task to both users and online commerce. More users are turning towards online shopping because it is relatively convenient, reliable, and fast; yet such users usually experience difficulty in probing for merchandises on the Web due to information overload. Online selling have often been stunned by the rich data they

have collected and find it challenging to endorse merchandise suitable to precise users. There is also the problem of futile consumption of the available huge amount of merchandise data from online transactions to support better decision making by both consumers and suppliers. To discourse these information overload problems, e-learning, e-commerce, e-newspapers data stores are now smearing mass customization ideologies not to the merchandises but to their staging in the online.

The Fuzzy Logic System could be understood as the nonlinear mapping of an input data set to a scalar output data set. Fuzzy sets obligate powerful decision making ability and hence attracted rising consideration and interest in modern information technology, production method, decision building, pattern acknowledgement, and diagnostics and data analysis among others. When a problem has vibrant or evolving behaviour, fuzzy logic is a suitable contrivance that deals with such problem. In short to say, fuzzy logic has metier in providing precise solutions to problems that encompass the manipulation of numerous variables.



The method of fuzzy logic systems are as follows:

- (i) Define input and output crisp variables
- (ii) Define the membership function
- (iii) Convert crisp input data into linguistic fuzzy values, using membership function, called fuzzification
- (iv) Evaluate the rules, using inference engine
- (v) Construct the output crisp data, from fuzzy linguistic values, called defuzzification.

Fuzzy logic bids several unique features that makes it a predominantly decent choice for many control problems.

i.) It is inherently robust since it does not require precise, noise-free inputs and could be programmed to fail safely if a feedback sensor quits or is destroyed. The output regulator is a smooth control function not withstanding an extensive assortment of input variations.

ii.) Since the Fuzzy logic checker processes the user-defined rules prevailing the target control system, it can be altered and tweaked easily to improve or radically alters system performance. New sensors can straightforwardly be fused into the system merely by engendering apt governing rules.

iii.) Fuzzy logic is not restricted to a few feedback inputs and control outputs, nor is it essential to measure or calculate rate-of-change restrictions in order to implement. This permits the sensors to be economical and imprecise thus keeping the inclusive system cost and intricacy low.

iv) Due to the rule-based process, any equitable number of inputs can be administered and numerous outputs engendered.

v) Create Fuzzy logic membership functions that express the implication (values) of Input / Output relationships used in the rules.

Fuzzification is conversion of crisp variables into linguistic variables, and it is the central unit for fuzzy logic system. Variable pertaining to linguistic sense such as age might obligate a value such as 'young' or 'old'. However, the noteworthy efficacy of linguistic variables is that they can be amended via linguistic verges pragmatic to primary terms. In 1973, Professor Lotfi Zadeh recommended the notion of linguistic or "fuzzy" variables. While variables in arithmetic typically gross numerical values, in fuzzy logic applications, the non-numeric linguistic variables are often used to simplify the countenance of rules and facts. The usage of linguistic variables in numerous applications cuts the overall computation complexity of the application. Linguistic variables obligate to be predominantly useful in complex non-linear applications.

### Proposed and Implemented system:

Here in this section we are giving comprehensive emphasis on the design of the system. Each and every stage of the offered system is well narrated here. Along with the elucidation the complete system is well presented using the architecture.

The complete system is dissected in four steps as discussed below.

#### 1. Pre-processing.

Pre-processing is vital step in data mining applications as it condenses the scope of the data required for processing. This condensed size minimizes the cost and space complexity of the system as fewer quantities of data are needed to be processed. Generally pre-processing encompasses of three steps.

- Special symbol removal.

Here all the special symbols from the content are removed e.g. !, @, #, \$, % etc. this special symbols are not going to contribute in result generation, hence its worth to remove all those symbols.

- Stop words removal.

Stop words are the words used as a supporting word in content to bring the semantics in the sentence. However after this discarding the meaning of the sentence is not changing too much extent. Hence they are removed here by maintaining one repository for comparison. This repository contains the 500+ stop words.

- Stemming.

Stems are used to derive word. Generally the words are derived to make the proper use of tenses. Unnecessarily this stems increase the system costing hence they are removed over here. No stemming algorithm is there which gives 100% accuracy.

#### 2. Feature Extraction.

As data contains tons of features it's not worth to consider the complete content for the further operations. Feature extraction is essential step in data mining. It is used to extract the required data i.e. features from the huge set of data. Here in our proposed work four features are extracted.

- Title sentence.

Title sentences are the one which represents the first sentence of the file content. The reason behind this extraction is to give a proper name to the cluster. Because each cluster is named by the title sentences.

- Numeric data.

Numeric data plays important role in file content as the most of the important data are represented using numerical values only. So by considering this thing we extracted numerical values from the file content.

- Proper nouns.

Proper nouns are the words which represent the person or place. For extraction of this feature a dictionary is used. So to access this dictionary jxlapi provides all the necessary functionalities.

- Top words.

Top words are the important words of the sentence. Here in this feature the frequency of the each word are found out. The word which repeat more time is needed to consider as it have the more weightage in the file content.

### 3. Master matrix creation.

Here in this step all the extracted features are taken as an input. From these entire features a one matrix is created. So particular feature of each file is compared with the respective feature of the other file. In this way all the four features are compared with four features of other file. This comparison led to a score of each file with other file.

Matrix creation process:

A weighted matrix is built, feature values are calculated against the every document in following way:

		D 1	D 2	D 3
		(T,N,P,Tw)	(T,N,P,Tw)	(T,N,P,Tw)
D 1	(T,N,P,Tw)	0		
D 2	(T,N,P,Tw)		0	
D 3	(T,N,P,Tw)			0
D n	(T,N,P,Tw)			

### 4. Fuzzy logic.

The generated score from matrix is taken as input. The smallest and biggest score is calculated. Exactly five ranges are calculated starting from smallest value and end to largest value. Now the score is assigned to each of the scores calculated in master matrix step by checking the occurrence of the score in these five ranges. Once score is calculate a threshold of 2 is set. The file having threshold more than 2 is added to cluster and discards the file which fails to satisfy the condition.

### Results and discussions:

To show the effectiveness of the system on experiment is conducted on java 1.6 based machine using Netbeans as an IDE on windows machine having 2GB rom and 500GB

HDD. After doing the experiment by providing the files from different categories such as text, pdf, and doc the following observation is led.

Number of documents	Time (seconds) to generate overlapping clusters
5	30
10	34
15	41
20	47
50	117
100	209
500	1004

Table 1: Time required against numbers of documents

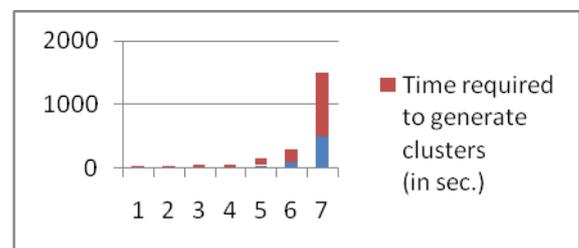
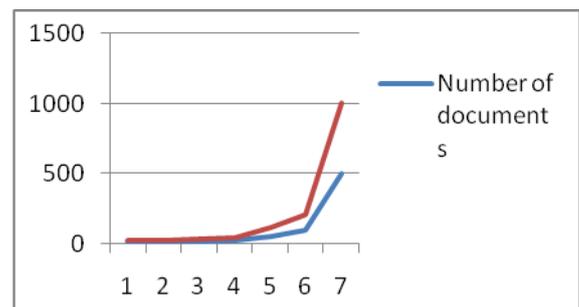


Figure 1: Performance measurement



The graph above signifies the clustering time. From the graph we can conclude that as the numbers of documents increase exponentially the required time to generate the clusters also increases in folds.

### References:

- [1] Kotsiantis, S. B., D. Kanellopoulos, and P. E. Pintelas. "Data pre-processing for supervised learning." International Journal of Computer Science 1.2 (2006): 111-117.
- [2] The wall, Mike. "A web crawler design for data mining." Journal of Information Science 27.5 (2001): 319-325.
- [3] M. Ram swami and R. Bhaskaran, "A Study on Feature Selection".
- [4] T.L. Hu, and J.B. Sheu, "A fuzzy-based customer classification method for demand-responsive logistical distribution operation", Journal of Fuzzy Sets and Systems, Vol. 19, 2003.
- [5] Frakes, William B. "Stemming Algorithms." (1992): 131-160.
- [6] Agrawal, Rakesh, and RamakrishnaSrikant. "Fast algorithms for mining association rules." Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215.1994.

- [7] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From data mining to knowledge discovery in databases." *AI magazine* 17.3 (1996): 37.
- [8] Munk, Michal, Jozef Kapusta, and Peter Švec. "Data preprocessing evaluation for web log mining: reconstruction of activities of a web visitor." *Procedia Computer Science* 1.1 (2010): 2273-2280.
- [9] Khasawneh, Natheer, and Chien-Chung Chan. "Active user-based and ontology-based web log data pre-processing for web usage mining." *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, 2006.
- [10] JH Paik, Mandar Mitra, Swapan K. Parui, Kalervo Jarvelin, "GRAS: An effective and efficient stemming algorithm for information retrieval", published in *ACM Transaction on Information System (TOIS)*, Volume 29 Issue 4, December 2011, Chapter 19, page 20-24
- [11] M. Bacchin, N. Ferro, and M. Melucci 2005. "A probabilistic model for stemmer generation". *Inf. Process. Manage.* 41, 1, 121–137.
- [12] WB Frakes, 1992, "Stemming Algorithm in Information Retrieval Data Structures and Algorithm", Chapter 8, page 132-139.
- [13] A. K. Jain, M.N. Murthy, and P. J. Flynn 1999. "Data clustering": A review. *ACM Comput. Surv.* 31, 3, 264–323.
- [14] The wall, Mike. "A web crawler design for data mining." *Journal of Information Science* 27.5 (2001): 319-325.
- [15] Castillo, Carlos. "Effective web crawling." *ACM SIGIR Forum*. Vol. 39.No. 1.ACM, 2005.
- [16] Olston, Christopher, and Marc Najork. "Web crawling." *Foundations and Trends in Information Retrieval* 4.3 (2010): 175-246.
- [17] Liu, Jin-Hong, and Yu-Liang Lu. "Survey on topic-focused Web crawler." *Appl. Res. Comp* 24.2629 (2007): 25.
- [18] T.V.Mahendra,,N.Deepika,N.KesacaRao," Data Mining for High Performance Data Cloud using Association Rule Mining", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol2, Issue 1, January 2012.
- [19] T.SunilKumar,Dr.K.Suvarchala, "A Study: Web Data Mining Challenges and Application for Information Extraction", *IOSR Journal of Computer Engineering (IOSRJCE)*, Vol 7, Issue3, Nov-Dec 2012, pp 24-29.
- [20] Darshna Navadiya, Roshni Patel, "Web Content Mining Techniques-A Comprehensive Survey", *International Journal of Engineering Research & Technology (IJERT)*, Vol. 1 Issue 10, December- 2012, pp.1-6
- [21] Govind Murari Upadhyay, Kanika Dhingra, "Web Content Mining: Its Techniques and Uses", *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 3, Issue 11, Nov. 2013, pp.610-613
- [22] Sandhya, Mala Chaturvedi, Anita Shrotriya, "Graph Theoretic Techniques for Web Content Mining", *The International Journal Of Engineering And Science (IJES)*, Vol 2, Issue 7 July 2013, pp.35-41
- [23] Xiaoqing Zheng, Yiling Gu, Yinsheng Li, "Data Extraction from Web Pages Based on Structural Semantic Entropy", *International World Wide Web conference Committee (IW3C2)*, April 2012, pp.93-102
- [24] Gengxin Miao, Junichi Tatemura, Wang-pin Hsiung, Arsany Sawires, Louise E. Moser, "Extracting Data Records from the Web Using Tag Path Clustering", *International World Wide Web conference Committee (IW3C2)*, April, 2009, pp.981-990.
- [25] M. Ramaswami and R. Bhaskaran, "A Study on Feature Selection
- [26] Huan Liu, Hiroshi Motoda, Rudy Setiono and Zheng Zhao, "Feature Selection: An Ever Evolving Frontier in Data Mining", *Journal of Machine Learning Research*, volume 10, June 2010, Hyderabad, pp. 4-13.
- [27] Le Song, Alex Smola, Karsten M. Borgwardt, Justin Bedo, "Supervised Feature Selection via Dependence Estimation", *proceedings International conference of Machine Learning (ICML)*, June 2007, USA.
- [28] Seoung Bum Kim, Panaya Rattakorn, "Unsupervised feature selection using weighted principal components", *International journal of Expert Systems with Applications*, Volume 38 Issue 5, May, 2011, pp. 5704- 5710.
- [29] Asha Gowda Karegowda, M.A. Jayaram, A.S. Manjunath, "Feature Subset Selection Problem using Wrapper Approach in Supervised Learning", *International Journal of Computer Applications*, Volume 1, No. 7, 2010, pp.13-17, ISSN: 0975 – 8887.
- [30] Sven F. Crone, Nikolaos Kourentzes, "Feature Selection for time series prediction—A combined filter and wrapper approach for neural networks", *Journal of Neuro computing*, Volume 73, Issues 10-12, June-2010, pp. 1923-1936, ISSN: 0925-2312.
- [31] "Effective and Efficient Feature Selection for Large-scale Data Using Bayes' Theorem", *International Journal of Automation and Computing*, Volume 6, Issue 1, Feb 2009, pp. 62-71.
- [32] Peng Liu, Naijun Wu, Jiaxian Zhu, Junjie Yin, and Wei Zhang, "A Unified Strategy of Feature Selection", *The Second International Conference on Advanced Data Mining and Applications (ADML 2006)*, China, August 2006, pp. 457 – 464.
- [33] Khalid, Samina, Khalil Tehmina, Nasreen Shamila, "A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning", *IEEE Science and Information Conference*, 2014, pp. 372-378.
- [34] Sitanshu Sekhar Sahu, Ganapati Panda, Ramchandra Barik, "A Hybrid Method of Feature Extraction for Tumor Classification Using Microarray Gene Expression Data", *International Journal of Computer Science & Informatics*, Volume 1, Issue 1, 2011, pp. 22 -26
- [35] Tomasz Kajdanowicz, Przemysław Kazienko, Piotr Doslak, "Label-Dependent Feature Extraction in Social Networks for Node Classification", *Lecture notes in computer science (Springer)*, volume 6430, 2010, pp.89-102
- [36] L. Huang, L. Dai, Y. Wei, and M. Huang, "A personalized recommendation system based on multi-agent", *Proceedings of the 2nd International Conference on Studies in Computational Intelligence 2010*.
- [37] R. Burke, "Knowledge-based recommender systems", *Encyclopaedia of Library and Information Systems*, Pp. 32-69, 2000.
- [38] J. B. Schafer, J. Konstan, and J. Riedl, "Electronic commerce recommender applications", *Journal of Data Mining and Knowledge Discovery*, Vol.5, Pp. 115–152, 2001.

- [39] T.L. Hu, and J.B. Sheu, "A fuzzy-based customer classification method for demand-responsive logistical distribution operation", *Journal of Fuzzy Sets and Systems*, Vol. 19, 2003.
- [40] I. Forenbacher, D. Perakovic, and I. Jovovic, "Model for Classification and selection mobile terminal devices applying fuzzy logic", *International Conference on Intelligent Computing*, University of Zagreb, Russia
- [41] S-S.Weng, and M-J. Liu, "Personalized product recommendation in e-commerce", *Proceedings of the 2004 IEEE international conference on e-technology, e-commerce and e-service*, Pp. 413–420, 2004.
- [42] H. Ying, "A Fuzzy Systems Technology: A Brief Overview" IEE Press, 2000.
- [43] J. Mendel, "Fuzzy Logic Systems for Engineering", *A Tutorial Proceeding of IEEE*, Vol. 83, Issue 3, Pp. 345-377, 1995.
- [44] D. Dubois and H. Prade, "An introduction to fuzzy systems", *Clin. Chim. octa* 270, pp 3–29, 1998.
- [45] A. Ali, and N. Mehli, "A Fuzzy Expert System for Heart Disease Diagnosis", *Proceedings of the International Multi Conference of Engineers and Computer Scientists*, Vol. 1, Pp. 134-139, 2010.
- [46] G.J. Klir, T.A. Folger, *Fuzzy sets, uncertainty and information*, Prentice Hall International, Englewood Cliffs, NJ, 1988.
- [47] R.R. Yager, *Fuzzy logic methods in recommender systems*, *Fuzzy Sets Syst.* 136 (2003) 133–149
- [48] R.R. Yager, F. Petry, *A framework for linguistic relevance feedback in content-based image retrieval using fuzzy logic*, *Inform. Sci.* 173 (4) (2005) 337–352
- [49] Guo, X., Lu, J.: *Intelligent E-Government Services with Personalized Recommendation Techniques*. *International Journal of Intelligent Systems* 22, 401–417 (2007)
- [50] Bezdec, J.C.: *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York (1981)
- [51] Gavalas, C. Konstantopoulos, K. Mastakas, and G. Pantziou, "Mobile Recommender Systems in Tourism," *Journal of Network and Computer Applications*, 39, 2014, pp.319-333
- [52] O. Khalid, M. Khan, S. U. Khan, and A. Zomaya, "Omni Suggest: A Ubiquitous Cloud based Context Aware Recommendation System for Mobile Social Networks," *IEEE Transactions on Services Computing*, vol. 7, no. 3, pp. 401-414, 2014
- [53] M. Deshpande, and G. Karypis, "Item-based top-n recommendation algorithms," *ACM Transactions on Information Systems*, 22, no.1, 2004, pp. 143–177.
- [54] P. H. Chou, P.H., Li, K. K. Chen, K.-K., and M. J. Wu, "Integrating web mining and neural network for personalized e-commerce automatic service," *Expert Systems with Applications* 37, no. 4, 2010, pp. 2898–2910.
- [55] Airasian, P. W., & Gay, L. R. (2003). *Educational research: Competencies analysis and application*. Englewood cliffs, N. J.: Prentice-Hall.
- [56] Bai, S. M., & Chen, S. M. (2008). *Automatically constructing concept maps based on fuzzy rules for adapting learning systems*. *Expert Systems with Applications*, 35(1-2), 41-49.
- [57] Baker, F. B. (1992). *Item response theory: parameter estimation techniques*. New York: Marcel Dekker.
- [58] Chen, C. M., & Chung, C. J. (2008b). *Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle*. *Computers & Education*, 51(2), 624-645.
- [59] Carlsson, C., Fedrizzi, M., & Fuller, R. (2004). *Fuzzy Logic in Management*. Kluwer Academy Publisher.
- Ebbinghaus, H. (1885). *Über das Gedächtnis. Untersuchungen zur Experimentellen Psychologie*. Leipzig: Duncker & Humblot, Germany.
- [60] Essalmi, F., Ayed, L. J. B., Jemni, M., Kinshuk, & Graf, S. (2010). *A fully personalization strategy of E-learning scenarios*. *Computers in Human Behavior*, 26(4), 581-591.
- [61] Ferreira, A., & Atkinson J. (2009). *Designing a feedback component of an intelligent tutoring system for foreign language*. *Knowledge-Based Systems*, 22(7), 496-501.
- [62] Hajek, P. (2006). *What is mathematical fuzzy logic*. *Fuzzy Sets and Systems*, 157(5), 597-603.
- [63] Huang, Y., & Bian, L. (2009). *A Bayesian network and analytic hierarchy process based personalized recommendations for tourist attractions over the Internet*. *Expert Systems with Applications*, 36(1), 933-943.
- [64] Hsu, M. H. (2008). *A personalized English learning recommender system for ESL students*. *Expert Systems with Applications*, 34(1), 683-688.
- [65] S. Lawrence and C. L. Giles, *Searching the world wide web*, *Science*, vol. 280, no. 5360, pp. 98–100, 1998.
- [66] Yuefeng Li and NingZhong, "Mining Ontology for Automatically Acquiring Web User Information Needs," *IEEE Transactions On Knowledge And Data Engineering*, VOL. 18, NO. 4, APRIL 2006.
- [67] Yuefeng Li, Wanzhong Yang, YueXu, "Multi-Tier Granule Mining for Rep-resentations of Multidimensional Association Rules," 0-7695-2701-9/06 *Proceedings of the Sixth International Conference on Data Mining IEEE 2006*.
- [68] Shady Shehata Fakhri Karray Mohamed Kamel, "Enhancing Text Clus-tering using Concept-based Mining Model," 0-7695-2701-9/06 *Proceedings of the Sixth International Conference on Data Mining IEEE 2006*.
- [69] Wai Lam, Miguel Ruiz, and Padmini Srinivasan, "Automatic Text Catego-rization and Its Application to Text Retrieval," *IEEE Transactions On Knowl-edge And Data Engineering*, VOL. 11, NO. 6. NOVEMBER DECEMBER 1999.
- [70] Helena Ahonen, Oskari Heinonen, Mika Klemettinen and A. Inkeri Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digi-tal Document Collections," *IEEE 1998*.
- [71] Sheng-Tang Wu Yuefeng Li YueXu Binh Pham Phoebe Chen, "Automatic Pattern-Taxonomy Extraction for Web Mining," *Proceedings of the IEEE/WIC/ ACM 0-7695-2100-2/04 International Conference on Web Intelligence*.
- [72] A Fuzzy, Incremental, Hierarchical Approach of Clustering Huge Collections of Web Documents *Int'l Conf. Internet Computing and Big Data COMP'13*
- [73] *Web Document Clustering Using Fuzzy Equivalence Relations -Journal of Emerging Trends in Computing and Information Sciences*
- [74] *Retrieval of Web Documents Using a Fuzzy Hierarchical Clustering - International Journal of Computer Applications (0975 – 8887) Volume 5– No.6, August 2010*