Document Clustering: TF-IDF approach

Prafulla Bafna
SICSR
Symbiosis International University
Pune, Maharastra , India

Dhanya Pramod
SCIT
Symbiosis International University
Pune, Maharastra , India

Anagha Vaidya
SICSR
Symbiosis International University
Pune, Maharastra, India

***Abstract*— Recent advances in computer and technology resulted into ever increasing set of documents. The need is to classify the set of documents according to the type. Laying related documents together is expedient for decision making. Researchers who perform interdisciplinary research acquire repositories on different topics. Classifying the repositories according to the topic is a real need to analyze the research papers.**

Experiments are tried on different real and artificial datasets such as NEWS 20, Reuters, emails, research papers on different topics. Term Frequency-Inverse Document Frequency algorithm is used along with fuzzy K-means and hierarchical algorithm. Initially experiment is being carried out on small dataset and performed cluster analysis. The best algorithm is applied on the extended dataset. Along with different clusters of the related documents the resulted silhouette coefficient, entropy and F-measure trend are presented to show algorithm behavior for each data set.

Keywords: TF-IDF, entropy, silhouette coefficient, hierarchical, fuzzy k-means, clustering

I. Introduction

Managing growing repositories of unstructured or semi structured documents in an organization is becoming increasingly difficult. The size and number of online and offline documents is increasing exponentially. The need for identifying groups of similar documents has also increased for either getting rid of multiple versions of same documents or extracting relevant set of documents from huge document repositories. It benefits many applications such as finding near duplicated web pages, replicated Web collections, detecting plagiarism etc. Web search engines are highly benefited as it can be used for focused Crawling. Forming group of documents is not the only challenge, but there is need to identify the relevant group for a newly arrived document. It can be achieved through feature selection mechanism. It means that if the features of newly arrived document can be identified and matched with the feature set for each group of documents from the existing corpus. The new document gets placed in its relevant group depending on the match found.

Clustering techniques can be applied only on the structured data. So unstructured data need to be converted to structured data. It expects to extract the terms from the documents, so documents represent rows and terms are placed in columns. The terms are in large number which causes the problem of dimension curse and decreases algorithm efficiency.

To reduce these terms some feature selection technique should be used. TF-IDF technique is used which eliminates the most common terms and extracts only most relevant terms from the corpus. [1]

We follow an approach in which we preprocess the corpus to remove noisy and less useful data. We apply TF-IDF followed by hierarchical agglomerative clustering and fuzzy K-means algorithm with iterations. We use silhouette coefficient to determine number of clusters. [34]

The paper is organized as follows. The background presents the relevant work of other researchers on the topic. The next section presents our approach. The experimental set up and results obtained are presented next. The paper ends with conclusion and future directions.

Background

To get the information of interest quickly, the plethora of documents accumulated in an organization from internet, digital libraries, medical documents, news, scientific document etc. need to be kept in organized manner[2]. Many applications benefit from grouping of the document sharing the common facts such as finding replicated Web collections, detecting plagiarism, Web search engines, identifying near duplicate Web pages, focused Crawling, identifying spams, grouping search engine results etc. By building taxonomy of documents the quality and diversity of query results can be improved[3,4].Some enterprise search engines do this automatically but scalability is the big issue as, no of documents are ever increasing in size and number. One need to create clusters of documents dynamically without disturbing existing group of documents. Dynamic document clustering reduces efforts and time as it processes the new document and assigns it to the relevant cluster directly without need of running entire algorithm all over again. So features of newly arrived document can be identified and those can be mapped to the features of existing clusters of documents. So in the dynamic document repository newly arrived document can be inserted into a relevant group without re-clustering[5]. There are several techniques of Feature Selection(FS) such as wrapper, filter, ensemble, hybrid etc. [6, 7, 8, 9, 10]. Clustering is one of the Filter based FS techniques.

Clustering is a widely known data mining technique and several algorithms have been proposed by different researchers [11,12, 13,14] . These methods are broadly classified as distance based, partition based, density based etc. According to the outcome of clustering methods, the methods are categorized as nonhierarchical or hierarchical. Non-hierarchical divides data set of N items into M clusters. Hierarchical clustering produces nested data set in which pairs

978-1-4673-9939-5/16/$31.00 ©2016 IEEE

of items or clusters are connected successively. Above methods are crisp methods in which one element can belong to just one cluster. But in probability based cluster, one element can belong to different clusters. Membership value of the element which belongs to more than one cluster is different and is associated with the particular cluster. This value lies between zero and one. However, the hierarchical methods are appropriate for cluster-based document retrieval. The commonly used hierarchical methods, such as single link, complete link, group average link, and Ward's method, have high space and time requirements. In order to cluster the large data sets with high dimensionality there is need to have a better algorithm such as the minimal spanning tree algorithms for the single link method, the Voorhees algorithm [15] for group average link, and the reciprocal nearest neighbor algorithm for Ward's method. Edie in[15] has stated that documents or terms both can be clustered, he also listed steps of clustering including selecting of the attributes on which items are to be clustered, selecting appropriate clustering method, creating the clusters or cluster hierarchies, interpreting clusters and validating the results etc.

Researchers have listed [16,15] different similarity measures and different types of hierarchical agglomerative clustering(HAC) as single, complete and average link and ward’s method. In case when the document collection is dynamic means new items being added or old ones are removed, one needs some mechanism for updating the cluster structure. HAC can be used for this purpose. A Query can be matched against the documents to retrieve document from the cluster. Two approaches can be used. A top-down search involves entering the tree at the root and matching the query against the cluster at each node, moving down the tree following the path of greater similarity. The search is terminated according to some criterion. A bottom-up search begins with some document or cluster at the base of the tree and moves up until the retrieval criterion is satisfied. Nearest neighbor clusters is a simple retrieval mechanism, i.e. retrieving two most similar documents. Choosing of clustering algorithm depends on application and parameters associated with it. But all these methods can be applied on structured data.

The data is classified into two main types structured and unstructured. Unstructured refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner[17,18]. As it contains dates, numbers as well as facts, it results in irregularities and ambiguities so it can’t be stored in fielded form of a database. Organizations accumulate huge unstructured data like call center logs, emails, documents on web, tweets, comments, blogs, customer reviews on product. etc. and the information is basically stored in the form of documents or files. There is need to convert unstructured data into structured one in order to extract information for some decision making. Chakraborty [19] specifies how to organize and analyze text data using SAS Text miner and sentiment analysis studio. Machine learning algorithms can process the data if the structure is specifically in rows and columns but unstructured data suffers from problems such as misspellings, short forms, acronyms, colloquialism, grammatical complexities, mixing of one or more languages etc. So It becomes difficult to analyze unstructured data in the same way as structured data. SAS® Text Analytics tools provide text analysis capabilities. SAS can be used for automatic routing, automatic content categorization, root cause analysis, trend analysis, text rule building etc. [20]. The other method is converting unstructured data into structured form by creating a Term Document Matrix. Document is viewed as a concept space and gets a tabular representation in the form of major attributes. This document representation in the form of major terms present is also called as vector space model(VSM). It is the model where each term represents the dimension of the document and the value represents the number of occurrences in the particular document [21].

To compare two documents, different similarity measures are considered by different clustering algorithms. Most commonly used similarity measures are: (i) Euclidean, (ii) Cosine, (iii) Pearson correlation and (iv) Extended Jaccard[Muhammad Rafi].The correlated concept based method and Testor theory are used to cluster documents. It discover topics based on topic names, occurrences and associations [22,12]. Topic detection is also the technique which detects relevant labels for the document clusters. This technique helps the user to navigate and retrieve the needed information quickly and efficiently.Cosine similarity measure is widely used because it gives the best result with less efforts.[23]. But challenge here is dimension curse because documents are made of high dimensional text and take long time to process . The high dimension can be reduced by retaining only important terms thus features extraction has to be done so that the best features are selected for clustering process. Preprocessing is done by removing noisy data that can affect the clustering results. Stop words removal, stemming and sentence boundary determination are performed during this stage [11, 24] .One of the widely used weighting schemes like TF-IDF (Term Frequency –Inverse Document Frequency) is used to convert a document into structured format. It is a numerical to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. The TF-IDF value increases proportionally to the number of times a word appears in the document, but is offset by the frequency of the word in the corpus, which helps to control for the fact that some words are generally more common than others. [23].

But an ideal document consistently makes use of synonyms for a single word so that same words generally do not repeat. Thus TF-IDF would not consider that word as significant with respect to that document and thus though important the word gets ignored. One way to solve this problem is inclusion of synonyms and finding out hidden relationship between the documents. Jayaraj Jayabharathy, G.Bharathi in [5,31] have given thought, in which synonyms are included and existing term set prepared by TF-IDF are enriched. This gives slightly improved results on some data sets.

There are some performance metrics like Variance, Entropy, F-measure and Purity, silhouette width etc. that are used to evaluate the quality of document clustering. F-measure combines the Precision and Recall from information retrieval process[2]. Generally cluster quality is measured using silhouette coefficient, F-measure, entropy and precision. The average silhouette width provides an evaluation of clustering validity, and might be used to select an ‘appropriate’ number of clusters[34] Entropy measures the uniformity or purity of a cluster and Precision directly reflects the performance of clustering .Entropy is used with various preprocessing methods such as wrapper, filter for feature elimination, reduction and selection [26, 27, 28, 30, 32, 8, 12].

.

II. Multistep algorithm

We use TF-IDF on a set of documents. Then cosine distance matrix is constructed and hierarchical agglomerative clustering and fuzzy K-means algorithm is applied to get the clusters. Following steps are applied

1. Preprocess the documents by removing stop words, stemming list etc.
2. Formulate term document matrix by applying TF-IDF and calculate cosine distance matrix.
3. Apply hierarchical agglomerative clustering and fuzzy k-means get the clusters.
4. Validate the quality of Clusters by calculating entropy for both algorithms.
5. Finalize the algorithm by looking into entropy and F-measure value
6. Use silhouette coefficient to finalize the number of clusters.
7. Apply finalized algorithm on extended dataset

The experiments carried out on different datasets presented in next section reflects that our approach gives better clusters of documents.

III. Experimental Set up

The experimental setup of multistep and fuzzy k-means algorithm consists of several existing tools. Such as Refine, R etc.

Initially we used artificially created data sets for checking efficacy of the algorithm. It comprises of different research papers related to feature selection. The domain knowledge of the researcher could be used to verify the results. Satisfied with the obtained results, we moved on to existing available datasets such as Emails, News 20, Reuters.

The steps and the packages used for each step is explained below

1. We used refine tool [31] to convert pdf/doc files to excel.
2. We use R programming tool to merge all rows into a single row.(library(XLConnect))
3. Same tool is used to formulate Term Document matrix after Appling TF-IDF (library(SnowballC), library(tm)
4. The outcome of the R program is ported to excel, so that it can be further processed. This TDM (StringR) in Excel is converted to cosine matrix as input to the hierarchical clustering algorithm and fuzzy k-means algorithm in R(library(hclust)). It gives the required dendrogram and clusters
5. Entropy and F-measure is calculated for each of the algorithm to decide best suitable algorithm.
6. The selected algorithm is applied on large datasets.

A. *Experiment 1*

TF-IDF algorithm was executed on 45 documents which are Research Papers on feature selection, News20, Reuters, Emails.The dendrogram with four clusters of 45 documents is presented in Graph2 and Graph3. The datasets are Reuters and E-mail resp.Entropy and F- measure was calculated for each dataset. Lowest entropy and highest F-Measure value indicates the Best algorithm. This algorithm is applied on the large dataset to get the clusters. The results are presented in Table1 and Table2 resp. It is clear that for New 20 and Reuters dataset, HAC is used on the large dataset. For Research Paper data Fuzzy K-means should be used on the large dataset. For E-mail dataset, either one of the algorithm can be applied ,as Entropy and precision values do not have a significant difference. Graph1 clearly states the difference in values of entropy and F-Measure obtained for two algorithms. Following nine figures shows the representative graphs on different datasets. Graph 7 clearly states that e-mails can be well classified with fuzzy k-means which produces non overlapping clusters .Graph 8 and 9 are representative examples of cluster determination using silhouette coefficient.

Table 1

		Fuzzy K-means		***HAC***		
Sr. no	***Data Set(45)***	***Entropy***	***F-Measure***	***Entropy***	***F-Measure***	***Number of clusters***
1	News 20	0.389763	0.785621	0.256213	0.861231	4
2	Reuters	0.421189	0.614523	0.112368	0.889023	6
3	Research Paper	0.245612	0.884532	0.678951	0.523410	4
4	Email	0.214561	0.754312	0.225641	0.745612	4

Graph 1

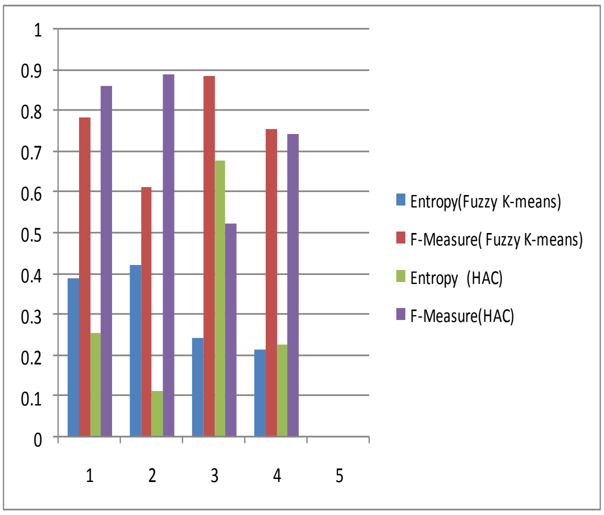


Table 2

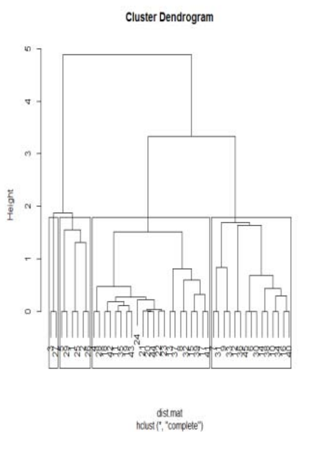
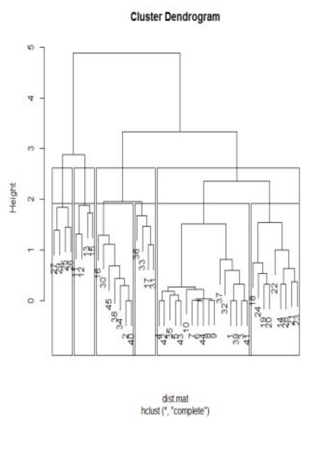
	Fuzzy K-means			HAC	
Data Set(810)	Entropy	F-Measure	Entropy	F-Measure	Number of clusters
News 20	-	-	0.356823	0.798967	4
Reuters	-	-	0.234167	0.801103	7
Research Paper	0.213189	0.798312	-	-	6
Email	0.289021	0.721390	0.361271	0.764321	6

B. *Experiment 2*

For this experiment one set of documents is selected from News20 dataset. We executed our algorithm from small to large sample from this data sets. It is very large dataset containing 10000 documents and classified into several groups. One set of documents belonging to a particular group (motorcycle,) were selected from NEWS20 dataset. The table shows total number of features and selected features.

Table 3

Dataset	Total terms	Selected terms
Religion	1543	32
Politics	1810	56
Baseball	1000	23
Hardware	1198	68
Motorcycle	456	55

Graph 2 : HAC for Reuter dataset (4 clusters)	Graph 3 : HAC for E-mail dataset (4 clusters)
	

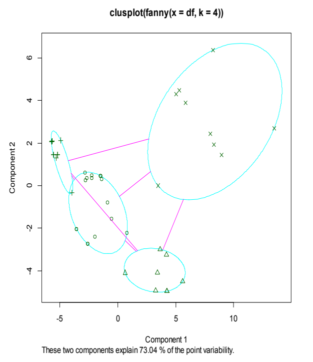
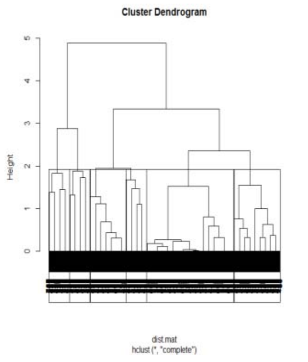
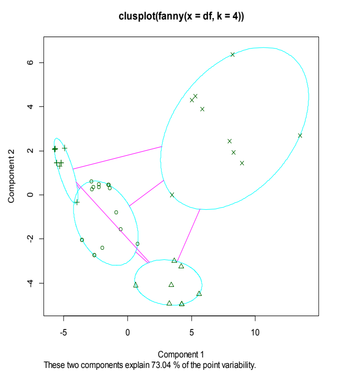
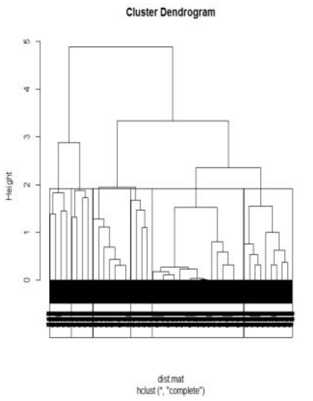
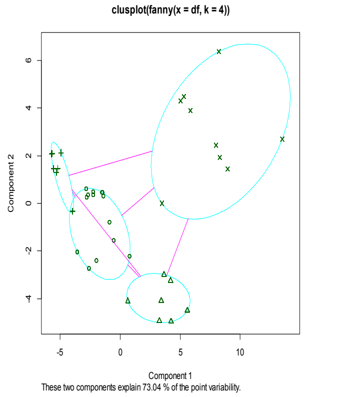
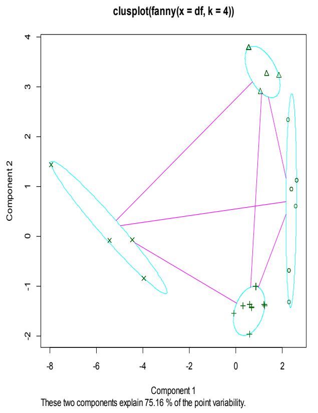
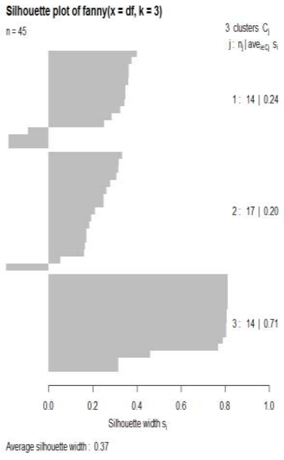
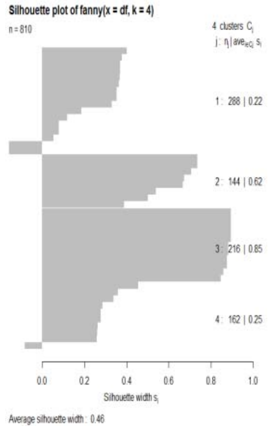
Graph 4 : fuzzy K-means plot for 45 Reuters (Overlapping clusters)	Graph : 5 Research Papers(810) 6 clusters
	

Figure1 fuzzy K-means plot for 45 News 20 documents : Overlapping clusters	Figure2 Research Papers(810) 6 clusters
	

Graph 6: fuzzy K-means plot for research papers (45)	Graph 7 Fuzzy K-means plot email dataset (810) (Non Overlapping clusters)
	

Graph 8 : Silhouette width for 45 research *papers*	*Graph 9* *Silhouette width for email dataset(810)*
	

IV. Conclusion

Our Experiment works in two phases the first phase detects the most suitable algorithm and in second phase ,it is applied to the extended data set .The results of the both phases are verified using different cluster analysis techniques. It can be used for wide ranges of problem varying from email sorting, research paper sorting .Identifying documents replicas existing in the corpus etc. Results obtained after processing various datasets shows efficiency of the algorithm. Future work is to use better semantic relativity concepts which might be domain specific but provide the better results.

References

[1] Ashish Moon T. Raju " A survey on document clustering with similarity measures, International Journal of Advanced Research in Computer Science and Software Engineering," Volume 3, Issue 11, pp. 599-601,November 2013.

[2] Joel W. Yu Jiao Reed, Thomas E. Potok, Brian TF-ICF "A new term weighting scheme for clustering dynamic data streams, International Conference on Machine Learning and Applications - CMLA ", pp. 258-263, 2006.

[3] Lavanya Pamulaparty, C.V. Guru Rao "A novel approach to perform document clustering using effectiveness and efficiency of simhash, International Journal of Engineering and Advanced Technology" Volume2, Issue3,pp.312-315, February 2013

[4] Yang Song "Boosting the feature space: text classfication for unstructured data on the web, Proceedings of the Sixth International Conference on Data Mining" 0-7695-2701-9/06.2006.

[5] Jayaraj Jayabharathy , Selvadurai Kanmani " Correlated concept based dynamic document clustering algorithms for newsgroups and scientific literature " springer decision analytics, 2014

[6] Peng-Fei-Zhu, Tian Hang Meng, Yun-long,Zhao, Rui-xan-ma, Qing-Hua Hu "Feature selection via mininmizing nearest neighbor classification error ,IEEE Proceedings of the Ninth International Conference on Machine Learning and Cybernetics" Volume .1,Page(s):506 - 511 ,11-14 July 2010.

[7] Tarek Amr "Survey on feature selection ,IEEE transactions on information forensics and security" Volume 3, no. 4, December , 2008

[8] Ashish Moon T. Raju " A survey on document clustering with similarity measures, International Journal of Advanced Research in Computer Science and Software Engineering" Volume 3, Issue 11, pp. 599-601,November 2013.

[9] Huan Liu Lei Yu "Towards integrating feature selection algorithms for classification and clustering" IEEE Transaction on Knowledge and Data Engineering, Volume 17 Issue 4, April 2005 ,Page 491-502.

[10] Yijun Sun, SinisaTodorovic, and Steve Goodison "Local learning based feature selection for high dimensional data analysis" IEEE Transaction on Pattern Analysis and machine Intillegence, Volume. X , 2010

[11] Damien Hanyurwimfura, Liao Bo, Dennis Njagi, Jean Paul Dukuzumuremyi "A centroid and relationship based clustering for organizing research apers,International Journal of Multimedia and Ubiquitous Engineering " Volume . 9 , No.3, Pp .219-234, 2014.

[12] Muhammad Rafi, Mohammad Shahid Shaikh ,"An improved semantic similarity measure for document clustering based on topic maps" 1303.4087 ,2013

[13] Shereen Albitar, Sebastien Fournier, Bernald Espinasse "An effective tf/idf-based text-to-text semantic similarity measure for text" spinger/chapter/10,WISE,2014,pp-105-114

[14] Qinbao Song.Jingjie Ni, Guangtao Wang "A fast clustering-based feature subset selection algorithm for high dimensional data" IEEE Transaction on Pattern Analysis and machine Intillegence, vol:25 no:1 ,2012.

[15] Edie Rasmussen "Clustering algorithms" University (http://orion.lcg.ufrj.br/Dr.Dobbs/books/book5/chap16.htm) Chapter 16: available on 03/03/2015

[16] B. Sindhuja, VeenaTrivedi "Usage of cosine similarity and term frequency count for textual document clustering," International Journal of Innovative Research in Computer Science & Technology (IJIRCST)ISSN: 2347-5552, Volume-2, Issue-5, pp. 9-12,September 2014

[17] Charles Elkan "Deriving tf-idf as a fisher kernel" springerSPIRE, LNCS 3772, pp. 296–301, 2005.,Springer-Verlag Berlin Heidelberg 2005.

[18] Jialei Wang, Peilin Zhao, Steven C.H. Hoi "Online feature selection and its applications " IEEE Transactions on knowledge and data engineering,Volume:26 , Issue: 3, pp. 698 - 710 ,2014

[19] Goutam Chakraborty, Murali Krishna "Analysis of unstructured data: applications of text analytics and sentiment mining" Paper 1288,SAS global forum,Washington,2014

[20] Padmapriya A recent survey on unstructured data to structured data in distributed data mining, International .Journal Computer Technology &Applications" Volume 5 (2),pp. 338-344,2014

[21] Muflikhah, L, Baharudin, B "Document clustering using concept space and cosine similarity measurement" International IEEE Conference Computer Technology and Development, 2009. ICCTD '09. Volume:1,09,PP 58-62

[22] Jayaraj Jayabharathy ,Selvadurai Kanmani ,N. Sivaranjan "Correlated concept based topic updation model for dynamic corpora" International Journal of Computer Applications (0975 – 8887) Volume 89 – No 10, March 2014

[23] ChiranjibiSitaula "Semantic text clustering using enhanced vector space model using nepali language" GESJ: Computer Science and pp 41-46 , 2012

[24] Muhammad Zubair Asghar, Aurangzeb Khan , Shakeel Ahmad ,Fazal Masud Kundi , A review of feature extraction in sentiment analysis, Journal of Basic and Applied Scientific Research" J. Basic. Appl. Sci. Res., 4(3), pp.181-186, 2014

[25] Julian Sedding, Dimitar Kazako "Wordnet-based text document clustering, Proceedings of the 3rd workshop on Methods in Analysis of Natural Language Data" pp 104-113, 2004.

[26] Zitao Liu " A feature selection method for document clustering based on part-of-speech and word co- occurrence" IEEE Conference on Fuzzy Systems and Knowledge Discovery, volume: 5 , pg 2331 - 2334 ,2010

[27] Mohammad-Amin Jashki "An iterative hybrid filter-wrapper approach to feature selection for document clustering" Proceedings of the 22nd Canadian Conference on Advances in Artificial Intelligence Volume 5549, 2009, pp 74-85.

[28] Huan Liu, Evolving "Feature selection, intelligent systems" IEEE (Volume:20 , Issue: 6) , pp 64 - 76 ,2005

[29] S.Sagar ,T.Sudha ,Novel "Feature selection method for classification of medical documents from pubmed" International Journal of Computer Applications (0975 8887),Volume 26,pp29-33, No.9,2011.

[30] Pabitra Mitra, C.A. Murthy, and Sankar K. Pal "Unsupervised feature selection using feature similarit", IEEE Transaction on Pattern Analysis

[31] G. Bharathi, D.Venkatesan "Improving information retrieval using document clusters and semantic synonym extraction" Machine Intelligence,pp-301-312 Issue: 3.Volume: 24 .

[32] Xiaofei He, Ming Ji, Chiyuan Zhang, and Hujun Bao "A variance minimization criterion to feature selection using laplacian regularizatio ",IEEE Transaction on Pattern Analysis and machine Intillegence, VOL. 33, NO. 10, 2011

[33] http://127.0.0.1:3333/ available on 03/03/2015

[34] Peter J. Rousseeuw, Journal of Computational and Applied Mathematics, Volume 20, November 1987, Pages 53–65

.