Dynamic Document Annotation for Efficient Data Retrieval

Deepali R Dagale
M.E Department of Computer Engineering
Indira College of Engineering and Management
Pune, India
deepali.dagale@gmail.com

Prof. PoornaShankar
Department of Computer Engineering
Indira College of Engineering and Management
Pune, India
poornashankar@indiraicem.aac.in

Abstract—Document annotation is considered as one of the most popular methods, where metadata present in document is used to search documents from a large text documents database. Few application domains such as scientific networks, blogs share information in a large amount is usually in unstructured data text documents. Manual annotation of each document becomes a tedious task. Annotations facilitate the task of finding the document topic and assist the reader to quickly overview and understand document. Dynamic document annotation provides a solution to such type of problems. Dynamic annotation of documents is generally considered as a semi-supervised learning task. The documents are dynamically assigned to one of a set of predefined classes based on the features extracted from their textual content. This paper proposes survey on Collaborative Adaptive Data sharing platform (CADS) for document annotation and use of query workload to direct the annotation process. A key novelty of CADS is that it learns with time the most important data attributes of the application, and uses this knowledge to guide the data insertion and querying.

Keywords-*Annotation, dynamic, classification, CAD, semi-supervised, content value, query workload.*

*****

I. INTRODUCTION

Data mining is the process of discovering meaningful data from large amounts of data stored in repositories using different technologies and techniques. Today, organizations are generating huge and growing amounts of data in different formats and different databases. Data mining is the process of analysing data and finding useful information patterns, associations, or relationships from it. For better decision making, the large amount data collected from different resources require proper methods for extracting knowledge from the databases.

Annotations are comments, notes, explanations, or external remarks which can be added to a document. Annotations are metadata as they give additional information about data. The main purpose of annotation is to facilitate future querying and also to improve the quality of searching. Data management tools like Microsoft’s SharePoint and Google Base support fixed or predefined annotation of the shared data which causes problems like schema explosion and inefficient data annotation, which in turn lead to unsatisfactory analysis and search performances. Annotation also helps to rank the documents based on various approaches like user’s feedback, previous search results, etc. Many systems, though, do not even have the basic “attribute-value” annotation that would make a “pay-as-you go” querying feasible. Annotations that use “attribute-value” pairs require users to be more principled in their annotation efforts. Difficulties results in very basic annotations, that is often limited to simple keywords. Such simple annotations make the analysis and querying of the data cumbersome. Users often use plain keyword for searches, or basic annotation fields, such as “date of creation” and “document owner”. The key goal of CADS is to minimize the cost of annotated documents that can be useful in query formulation. Our key goal is to encourage the process of annotation of documents at time of creation.

In Collaborative Adaptive Data Sharing (CADS) use the query workload for annotation processing by scrutinizing the content in the database. Similar kind of system has been developed in the upcoming year to improve the effective data management. These applications have to concentrate on the process of information extraction from the document. Bayesian theorem is used to calculate frequency of keywords based on contend and query workload. The results can be either fully match or partially match. But the preference is given to prior one. The classification driven text documents retrieval framework based on filtering and similarity fusion by employing supervised learning techniques. Various algorithms are used to retrieve the data, Bayesian theorem is used to calculate frequency of keywords based on content and query workload. It should retrieve the result superior to other approaches. Based on content and query workload documents are retrieved and ranking scheme is used to prioritize relevant results for the particular query.

II. RELATED WORK

An increasing amount of information becomes available in the form of electronic documents. There is need to intelligently process such texts makes to understand depth of knowledge from text. The lacking depth of knowledge understanding methods such as information extraction (IE) is useful. The review is done to get the insight of the methods, their shortcomings which can be overcome by giving the domain specific information, as input and storing the domain specific keywords in database with meaningful text and giving the related news as output.

The objective of this design review is to study and analyze the work carried and published by different researchers and authors in the domain of Document clustering and classification.

Eduardo J. Ruiz, VagelisHristidis, and Panagiotis G. Ipeirotis, proposed a paper, "Facilitating Document Annotation Using Content and Querying Value". In this paper, they proposed CADS (Collaborative Adaptive Data Sharing) platform which is an "annotate-as-you-create" infrastructure that facilitates fielded data annotation. The main contribution of this paper is to use the query workload directly to direct the annotation process, in addition to examining the content of the document. In other words, they are trying to prioritize the annotation of documents towards generating attribute values for attributes that are often used by users to query the database[1].

MohdNaghibzadeh (2010), proposed XCLS (XML Clustering by Level Structure) algorithm to cluster XML documents by considering their structures. This algorithm represents XML structure in a new way called level structure. Level structure only uses the elements or tags of the XML document and ignores their contents and attributes. Using a global criterion for computing similarity is a considerable point in incremental methods[2].

SatishMuppidi, MohdNaghibzadeh (2015), proposed the document clustering with Map/Reduce using Hadoop framework. The algorithm is to sort data set and to convert it to (key, value) pair to fit with Map/Reduce. The operations of distributing, aggregating, and reducing data in Map/Reduce should cause the bottle-necks. MapReduce is a striking framework because it allows users to decompose the data involved in computing documents similarity into multiplication and summation stages separately in a way that it is well matched to effective disk access across several nodes[3].

Habibi, M., Popescu-Belis, A(2015), addresses the problem of keyword extraction from conversations, with the goal of using these keywords to retrieve, for each short conversation fragment, a small number of potentially relevant documents, which can be recommended to participants[4]. A method is used to derive multiple topically separated queries from this keyword set, and clustering technique is used to divide the set of keywords into smaller topically independent subsets constituting implicit queries in order to maximize the chances of making at least one relevant recommendation when using these queries to search over the English Wikipedia.

LeventBolelli, SeydaErtekin(2007), proposed K-SVMeans, a clustering algorithm for multi-type interrelated datasets that integrates the well-known k-means clustering with the highly popular Support Vector Machines. The experimental results on authorship analysis of two real world web-based datasets show that K-SVMeans can successfully discover topical clusters of documents and achieve better clustering solutions than homogeneous data clustering[5].

Y.SureshBabu, K.VenkatMutyalu(2012), presents fast greedy k-means algorithm for document clustering as it is very much accurate and efficient and also introduce an efficient method to calculate the distortion for this algorithm. This helps the users to find the relevant documents more easily than by relevance ranking[6].

V.M.Navaneethakumar, Dr.C.Chandrasekar(2012), proposed a Conceptual Rule Mining On Text clusters to evaluate the more related and influential sentences contributing to the document topic. In this paper, the conceptual text clustering extends to web documents, containing various markup language formats associate+ed with the documents (term extraction mode). Based on the markup languages like presentations, procedural and descriptive markup, the web document's text clustering is done efficiently using the concept-based mining model. Experiments are conducted with the web documents extracted from the research repositories to evaluate the efficiency of the proposed consistent web document's text clustering using conceptual rule mining with an existing An Efficient Concept-Based Mining Model for Enhancing Text Clustering[7].

Mr. Sumit D. Mahalle, Dr. KetanShah(2012), proposed Semantic based approach for document clustering which is mainly based on semantic notations of text in documents. In the Semantic document clustering system parse the web documents into two way, first is syntactically and second is semantically[8]. Syntactical parsing can ignore the less important data from documents to get proper data, which is to passed to next step. Then in next step i.e. Semantic parsing can apply on the parsed syntactic data which cluster the documents properly and give the needed response to user at the time of data mining which is not accurately in traditional methods. Then in next step these text files will be pass into semantic clustering, to get the membership value of each text file. So finally formed the clusters of text files which can be calculated by comparing its membership values with each other. "Document Clustering by using semantics" is a technique which is direct database, there are very few technique present which are based on textual data clustering. As searching space is small after clustering with semantic approach, system need very less time to search through billions of web pages or documents in fraction of seconds or less.

ChandanJadon, Ajay Khunteta(2013), proposed new approach of document representation, document - term matrix .The rows represent the documents and the columns represent the terms number. The terms are arranged in such a manner,the term number is first in the list whose weight is highest and they are arranged in decreasing order of weight(frequency in document). Then algorithm is used to form the clusters based on term frequency and the domains[9].

Jiashen Sun; Xiaojie Wang(2011), presents Web page clustering based on semantic or topic promises improved search and browsing on the web. Intuitively, tags from social bookmarking websites such as del.icio.us can be used as a complementary source to document thus improving clustering of web pages. They proposed a novel model which employs topic model to associate annotated document with a distribution of topics, and then constructs a graph including tags, document and topics by performing a Random Walks for clustering[10].

Baghel, R.; Dhir, R.(2010), proposed a novel technique of document clustering based on frequent concepts. The proposed FCDC (Frequent Concepts based Document Clustering), a clustering algorithm works with frequent concepts rather than frequent item sets used in traditional text mining techniques. Many well-known clustering algorithms deal with documents as bag of words while they ignore the important relationship between words like synonym relationship. The proposed algorithm utilizes the semantic relationship between words to create concepts. It exploits the WordNet ontology in turn to create low dimensional feature vector which allows developing a more accurate clustering[11].

Durga et al(2009), proposed an algorithm for clustering unstructured text documents using native Bayesian concept and shape-pattern matching. The Vector Space Model is used to represent their dataset as a term-weight matrix[12]. In any natural language, semantically linked terms tend to co-occur in documents. This information is used to build a term-cluster matrix where each term may belong to multiple clusters. The native Bayesian concept is used to uniquely assign each term to a single term-cluster. The documents are assigned to clusters using mean computations. They applied shape pattern-matching to group documents within the broad clusters obtained earlier.

Shankar Setty, RajendraJadit, SabyaShaikh, ChandanMattikalliand VmaMudenagudi proposed a paper “Classification of Facebook News Feeds and Sentiment Analysis.” In this paper, they presented a system for classification of facebook news feeds. They developed a model to classify posts appearing on usersfacebook wall to find most important news feeds. Also it helps to automatically detect the sentiments of the user. SVM classifier learner model is used for structuring the data on facebook[13].

MojganFarhoodi, and AlirezaYariproposed the paper “Applying machine learning algorithms for automatic Persian text classification.”Classification plays an important role in information retrieval. In this paper, they examined two efficient machine learning algorithm for Persian text document. Performance of KNN is better than SVM[14].

Kevin HsinYih Lin, Changhua Yang, and HsinHsi Chen proposed a paper, “Emotion Classification of Online News Articles from the Reader’s Perspective.” In this paper, they automatically classify documents into reader emotion categories. This paper mainly addresses the reader perspective. Helps user to retrieve relevant content from web search engine[15].

A.S.M Shihavuddin, Mir NahidulAmbia, and Mir Mohammad NazmulArdinproposed a paper, “Prediction of stock price analyzing the online financial news using Naive Bayes classifier and local economic trends”. In this paper, Naïve Bayes algorithm is used to predict the data of stock market. It was observed by author that performance of Naïve Bayes is based on frequency of data instances, attribute values and target values for accuracy prediction[16].

M. JanakiMeena , and K. R. Chandran proposed a paper,” Naive Bayes Text Classification with Positive Features Selected by Statistical Method.” In this paper, they proposed an algorithm CHIR which is statistical based approach. This algorithm improved the accuracy of most popular text classification algorithm i.e Naïve Bayes.[17]

III. PROPOSED SYSTEM

The proposed search engine for News Blog has two phases: annotation phase and search phase. Figure 1 shows the architecture of the proposed document annotation system. The system has three databases: document collection, query collection and annotation database. To create this system, eclipse IDE using Java for front end designing and MySQL Server as back end will be used.

The first step of annotation strategy is document uploading. The author or creator of the document can upload his document in to the repository. After uploading the document, system analyses the document text for generating the best attributes for annotation. The author can upload his document by clicking on the upload button. Then the author will submit the document for annotation insertion form generation. After the document uploading, system will analyses document collection as well as the query collection for generating the best attribute value pairs. For suggesting an attribute for annotation the attribute must satisfy two conditions: (1) attribute should have high querying score and (2) attribute should have high content score.

Query score is the frequency of occurrence of an attribute in the query collection. Content score is the frequency of occurrence of an attribute in the document collection. Based on the content score and query score the attributes will be suggested in the annotation insertion form. After generating the form, the author can check the values of the generated attributes. The author can also add his own attribute to the form and finally submit the form along with the document for storing in the system repository.

The second phase of the document annotation system is document searching. The user can search for documents using query form. The user can provide search conditions as attribute name and attribute value pairs for searching the documents. After submitting the form the system will search for documents satisfying the query conditions in the annotation database. Thus the system will save the search time in finding relevant documents in traditional systems. Search results will be a list of documents that satisfying the user query conditions. The user can see a list of documents and on clicking the document names he can view the contents of the document. If there is no document that satisfies the user query conditions, then there is no document to display and return the result as no match found. Search queries given by user will then store to query collection database for training the document annotation system in future annotations.

Score of an attribute in a document is found by using a probabilistic approach. Total score of an attribute is the product of content score and query score. Attributes with high score is more relevant to that document. Finally attributes will display in an annotation insertion form in the decreasing order of their score.

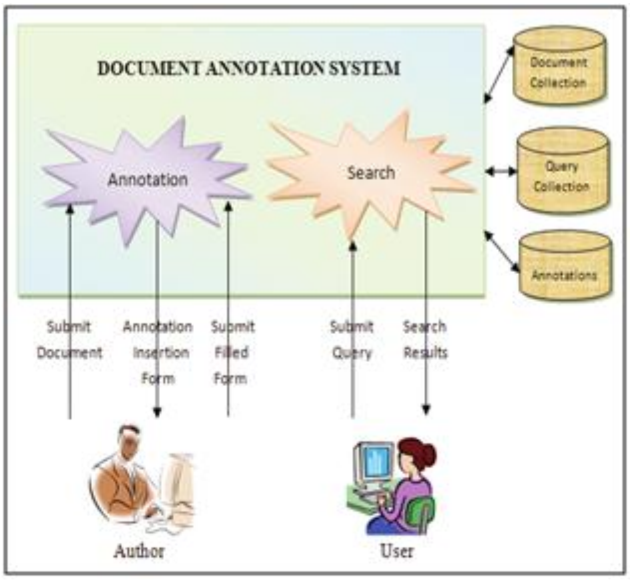


Figure 1: System Architecture

IV. MODULES DESCRIPTION

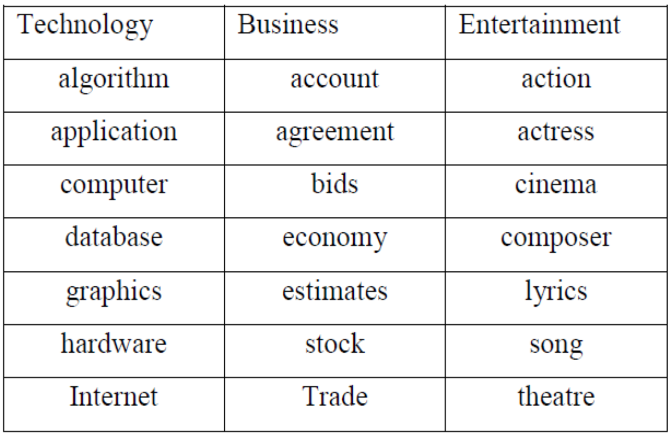
Dataset consists of google news(.txt files) from various domains, these files are preprocessed to remove stop words and thereafter stemming process is appied. Term frequency is calculated, after that keywords are clustered using k-mean algorithm. The user searches based on keywords, relevant documents are retrieved from respective domains.

MODULES

- Domain Specific Dictionary
- Input Selection
- Pre-processing
- Clustering
- Searching
- Performance Evaluation.

Domain Specific Dictionary

First, we have analysed news from various domains and maintained domain specific dictionary which contains keywords from each stream as shown below in figure 2. According to concept dictionary, three streams as Technology, Business and Entertainment are the domains.



Technology	Business	Entertainment
algorithm	account	action
application	agreement	actress
computer	bids	cinema
database	economy	composer
graphics	estimates	lyrics
hardware	stock	song
Internet	Trade	theatre

Figure 2: Domain Specific Dictionary

Input Selection

In this module, the input selection which is the initial step, is collection of google news from Technology, Business and Entertainment domain. Then these news is preprocess to remove stop words and stemming is applied.

Pre-processing

This module consist of stop words removal and stemming process.

1.Stop words Elimination

Stop words are words which are filtered out prior to, or after, processing of natural language data. A stop word is a commonly used word in our daily life, that a search engine has been programmed to ignore, both when searching and when retrieving them as a result of a search query. The major work is to identify the mostly weighted words are called as keywords for the documents that reduce the dimensions of the matrix. Stop word elimination is done based on ASCII values of each letter without considering the case (either lower case or upper case) and sum the each letter corresponding ASCII value for every word and generate the number. Assign number to corresponding word, and keep them in sorted order.

2. Stemming

Stemming is the process for reducing modified words to their stem, base or root forms generally a written word form. Stem is a part of a word like ing, er, etc., The term is used with slightly different meanings. An algorithm for removing derivations endings and inflectional in order to reduce word forms to a common stem. In this stemming algorithm the suffixes and prefixes were eliminated according to the conditions by which the stemming procedure was applied.

A stemmer for English, for example, should identify the string "cats" (and possibly "catlike", "catty" etc.) as based on the root "cat", and "stemmer", "stemming", "stemmed" as based on "stem". A stemming algorithm reduces the words "fishing", "fished", and "fisher" to the root word, "fish". On the other hand, "argue", "argued", "argues", "arguing", and "argus" reduce to the stem "argu" (illustrating the case where the stem is not itself a word or root) but "argument" and "arguments" reduce to the stem "argument".

Clustering

Clustering is a group of similar objects. Cluster is a collection of data objects, that are contains all similar objects, and these objects are dissimilar to the other clusters objects. A cluster of data objects can be treated collectively as one group and so may be considered as a form of data compression. Objects within the cluster have the high similarity in comparison to one another but are dissimilar to objects in the other clusters. Clustering principle:" Maximizing the intra class similarity and minimizing the inter class similarity". In this module , Document clustering uses domain specific dictionary to cluster the documents. Domain specific dictionary contains sets of words that describe the contents within the cluster. k-mean algorithm is used for clustering. For the k-mean algorithm we have to decide value of k when beginning algorithm starts, it is noticed that different value of k will cause different levels of accuracy of the grouping.

Searching

When user search for document, based on our user query system retrieves the documents, the goal of our system is to maximize the number of relevant documents in the ranked list as well as making sure that they are high up in the ranked list. Additionally we provide annotations to every search that is been in query module.

Performance Evaluation

In this module, the evaluation which done by graphical representation, it shows simple search and cluster search approach for user query. It also shows time efficiency between two approach graphically.

The architecture diagram represents the overall structure of the system. The documents clustered using k-mean algorithm. The dataset is collected from googlenews, it is collection of text files, given as input to the system.

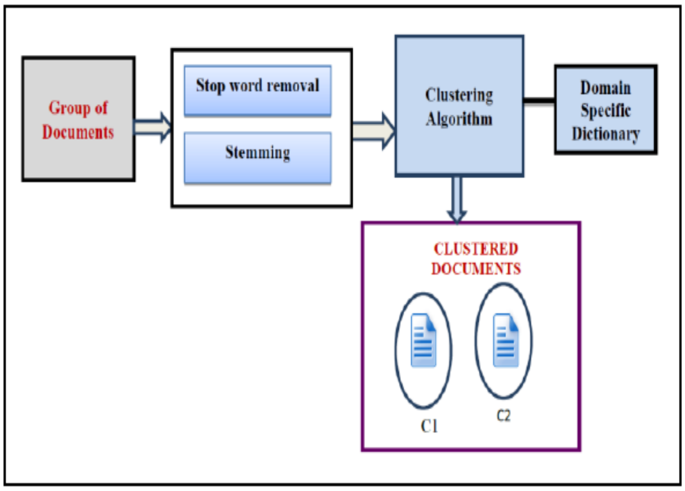


Figure 3: Architecture of Document Clustering

V. Block Diagram

The following figure 4 represents the block diagram for proposed system which shows how the system automatically suggest attribute for annotation. The User/ Publisher upload the text documents. Extract each and every words present in the document. Then it performs stemming on document text. Adds words obtained after stemming to list L(say). Then Add words from list L which appears in annotation database to list S. In next step, it find score of words in list S. Score of an attribute in a document is found by using a probabilistic approach. Total score of an attribute is the product of content score and query score. Attributes with high score is more relevant to that document. Search results will be a list of documents that satisfying the user query conditions. The user can see a list of documents and on clicking the document names he can view the contents of the document. If there is no document that satisfies the user query conditions, then there is no document to display and return the result as no match found.

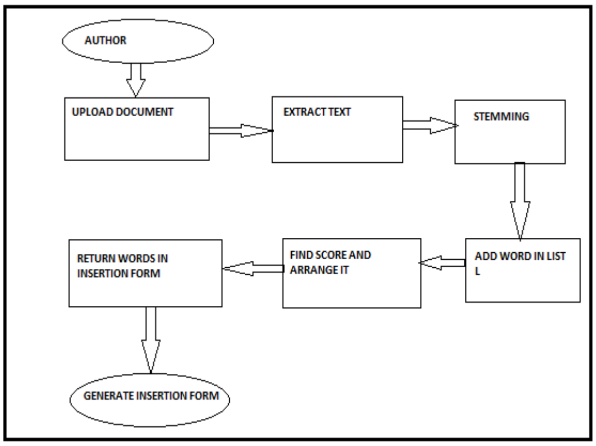


Figure 4: Block Diagram for Document Annotation System

VI. KEY FEATURES:

1. Statistical Learning Method

Bayesian Learning is used for Attribute Suggestion. Rather than choosing the most likely model or delineating the set of all models that are consistent with the training data, another approach is to compute the posterior probability of each model given the training examples. The idea of **Bayesian learning** is to compute the posterior probability distribution of the target features of a new example conditioned on its input features and all of the training examples.

2. **Bernoulli Model or Bayes theorem** is used for calculating the score of attributes.

In probability theory and statistics, **Bayes' theorem** (alternatively **Bayes' law** or Bayes' rule) describes the probability of an event, based on conditions that might be related to the event. For example, if cancer is related to age, then, using Bayes' theorem, a person's age can be used to more accurately assess the probability that they have cancer.

3. **Pipelined Algorithm** is used to compute the top-k attributes with highest score.

4. K-means Algorithm

K-means is the most important flat clustering algorithm. The objective function of K-means is to minimize the average squared distance of objects from their cluster centers, where a cluster center is defined as the mean or centroid μ of the objects in a cluster C.

The ideal cluster in K-means is a sphere with the centroid as its center of gravity. Ideally, the clusters should not overlap.

5. Stop words removal Algorithm

In computing, stop words are words which are filtered out prior to, or after, processing of natural language data (text). There is not one definite list of stop words which all tools use and such a filter is not always used. Some tools specifically avoid removing them to support phrase search.

6. Stemming Algorithm

The Porter stemming algorithm (or 'Porter stemmer') is a process for removing the commoner morphological and inflexional endings from words in English. Its main use is as part of a term normalisation process that is usually done when setting up Information Retrieval systems.

7. Threshold algorithm with Restricted Sorted Access (TAz):

Threshold algorithm with Restricted Sorted Access is used (TAz) for combining QV(Query value) and CV(Content Value).

VII. CONCLUSION

The proposed system will provide solution to annotate the document at time of uploading and also works on user's querying needs. The proposed architecture works on the content of document and also analyse the user queries. The query work load can greatly improve the annotation process and increase the utility of shared data. Bayesian approach is an effective technique to extract the information from the document. Along with annotation document, pattern mining is the technique that helps the user to map document with frequent pattern and use pattern at the time of searching. The annotation and pattern matching technique provides flexible and complete solution for document tagging and searching.

REFERENCES

[1] Eduardo J. Ruiz, Vagelis Hristidis, and Panagiotis G. Ipeirotis(2014), "Facilitating Document Annotation Using Content and Querying Value", IEEE TRANSACTIONS, VOL. 26, NO. 2.

[2] Mohd. Alishahi, Mohd Naghibzadeh (2010), "Tag Name Structure Based Clustering of XML Documents ", International Journal of Computer and Electrical Engineering, Vol. 2/1, 1793-8163.

[3] Satish Muppidi, Mohd Naghibzadeh (2015), "Document clustering with Map reduce using Hadoop framework", International Journal on recent and innovation Trends in Computing and communication, Vol. 3/1, 409-413

[4] Habibi, M.; Popescu-Belis, A.(2015),"Keyword Extraction and Clustering for Document Recommendation in Conversations" Volume: 23, Pages: 746 - 759

[5] Levent Bolelli, Seyda Ertekin(2007), "K-SVMeans: A Hybrid Clustering Algorithm for Multi-Type Interrelated Datasets",7695-3026-5/07

[6] Y.SureshBabu, K.Venkat Mutyalu(2012), "A Relevant Document Information ClusteringAlgorithm for Web Search Engine",International Journal of Advanced Research in Computer Engineering & Technology,Volume 1.

[7] V.M.Navaneethakumar, Dr.C.Chandrasekar(2012), " A Consistent Web Documents Based Text Clustering Using Concept Based Mining Model", International Journal of Computer Science Issues, Vol. 9

[8] Mr. Sumit D. Mahalle, Dr. Ketan Shah(2012), "Document Clustering by using Semantics", International Journal of Scientific & Engineering Research, Volume 3.

[9] Chandan Jadon, Ajay Khunteta(2013), "A New Approach of Document Clustering", International Journal of Advanced Research in Computer Science and Software Engineering,Volume 3.

[10] Jiashen Sun; Xiaojie Wang(2011),"Annotation-aware web clustering based on topic model and random walks" International Conference on Cloud Computing and Intelligence Systems (CCIS), Pages: 12 - 16.

[11] Baghel, R.; Dhir, R.(2010), "Text document clustering based on frequent concepts", International Conference on Parallel Distributed and Grid Computing (PDGC),Pages:366-371.

[12] Durga Toshniwal, Rishiraj Saha Roy(2009), Clustering Unstructured Text Documents Using Naive Bayesian Concept and Shape Pattern Matching, IJACT : International Journal of Advancements in Computing Technology, Vol. 1, No. 1, pp. 52 ~ 63.

[13] Shankar Setty, Rajendra Jadit, Sabya Shaikh, Chandan Mattikalli and Vma Mudenagudi* “Classification of Facebook News Feeds and Sentiment Analysis.”

[14] Mojgan Farhoodi, and Alireza Yari “Applying machine learning algorithms for automatic Persian text classification.”

[15] Kevin HsinYih Lin, Changhua Yang, and Hsin Hsi Chen, “Emotion Classification of Online News Articles from the Reader’s Perspective.”

[16] A.S.M Shihavuddin, Mir Nahidul Ambia, and Mir Mohammad Nazmul Ardin, “Prediction of stock price analyzing the online financial news using Naive Bayes classifier and local economic trends”.

[17] M. Janaki Meena , and K. R. Chandran,” Naive Bayes Text Classification with Positive Features Selected by Statistical Method.”