Enhancing Computer Inspection Using Document Clustering for Analysis

Mr.Chandrasekhar M S1, Mr.Subhash B N2, Mr.Satish V3, Mr.Bharath J4
1, 2, 3 UG Student, Dept of CSE, RRCE, 4Asst.Prof, Dept of CSE, RRCE, Bangalore-74
chandrashekarms214@gmail.com, subhashbn71@gmail.com, RRCE,sathishv138@gmail.com, bharath5911@gmail.com

ABSTRACT-In document analysis, Computers having huge amount of data files really creates disorder to analyze it, most of the data consist in those files will be unstructured whose analysis will be difficult. Therefore, we present an approach that reduces the effort of analysis by clustering the document. Clustering is a division of data into groups of similar objects. The clustering techniques used in our approach are k-means and incremental mining, both this algorithm facilitate in discovering new and useful information from the documents under analysis. Finally, we show our results in graphs for better summarization and visual presentation purpose.

Keywords- Document clustering, text mining, Analysis

I. INTRODUCTION

The most common style of unattended learning is by clustering and this is often the major distinction between clustering and classification. No super-vision implies that there's no human professional who has allotted documents to categories. The goal of a document cluster theme is to reduce intra-cluster distances between documents, whereas increasing inter-cluster distances (using AN applicable distance measure between documents). A distance measure (or, dually, similarity measure) therefore lies at the center of document clustering. The large form of documents makes it nearly impossible to form a general algorithmic rule which might work best just in case of all types of data

Many times, it takes a lot of time to scan all the computer information and look for the required file. Thus, a professional usually scans the computers manually and tries to gather needed data. But, it may take lots of efforts and time a protracted time. Thus, so as to beat this drawback, the idea of document cluster is terribly helpful. The clustering algorithms are often helpful wherever no information concerning the information in connected document square measure best-known a priori [2],[3]. Thus, clustering helps a lot to partition information into cluster of connected documents. There are numerous cluster approaches with well-known algorithms like k-means, k-medoid, single link, complete link, etc. [1]. In our proposed system, we've used k-means and incremental mining algorithm. K-means algorithms works on comparatively validity index to estimate the cluster numbers automatically and incremental mining algorithm uses cosine similarity measurement. We‘ve also targeted on preprocessing steps like removal of stop words and to stem the words which may facilitate to form the data which will be organized. Thus, smart preprocessing techniques will facilitate to scrub and create that data to be effectively utilized in cluster method. Our proposed methodology forms the most clusters wherever connected documents found. We tend to confer our results with the help of graph for higher summarization and visual presentation purpose.

II. LITERATURE SURVEY

There are studies regarding use of clustering algorithms in the field of text analysis of documents. Most of the studies describe the use of algorithms for clustering data e.g., k-means, k-medoid, Fuzzy C-means(FCM), single link, complete link Expectation-Maximization (EM) for unsupervised learning of Gaussian Mixture Models, and Self- Organizing Maps (SOM). SOM [2] is generally similar to K-means but are usually less efficient. SOM based algorithms were used for clustering files and making the decision-making process performed by the examiners more efficient and accurate. The files were clustered by taking into consideration their creation dates/times and their extensions.

Jian Ma, Wei Xu [8], proposes an Ontology based Text Mining (OTMM) method to cluster research proposals in a research funding agency. SOM is applied to cluster the research proposals on the basis of similarity. After the grouping, reviewers are assigned with the proposals. Hence this approach reduces the time of grouping the research proposals. Qiujun [12], proposes an approach for extraction of reports content using similarity measure based on edit distance to separate the news content from clattering data. This paper describes regarding the correct extraction of reports content from web pages. The algorithms used with this methodology are less advanced with high accuracy and efficiency rate. S. C. Punitha, M. Punithavalli[11], studied two approaches for text clustering and compared them. First method relies on pattern recognition with semantic driven ways for clustering text documents. Second method is ontology based mostly text clustering approach. Each algorithm are analyzed in terms of efficiency and speed of clustering, however the performance of ontology based approach was higher in terms clustering quality, but lack in speed.

III. PROPOSED SYSTEM

In our proposed system we have decided to choose two main clustering algorithm i.e. k-means and Incremental mining algorithm. These algorithms run with different combinations of their parameters, resulting in different algorithmic instantiations. The aim is to dynamically create the cluster, carry out the analysis and to measure the performance of used algorithms. The methodology is

to accept the data from user including large sets of text documents, apply the pre-processing on the text file then clustering it using clustering algorithms. The clustering is done with reference of sample document which will be match with all other documents and clusters are formed by analyzing the difference between documents and the centroids used in clustering algorithms.

IV. SYSTEM ARCHITECTURE

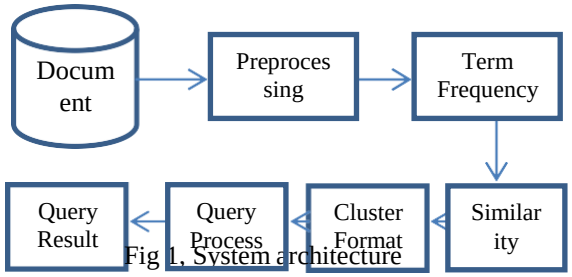

Fig 1, System architecture

Before providing documents to clustering algorithm, we tend to perform some preprocessing steps. Specially, Stop-words such as prepositions, pronouns, articles are removed. Then, we tend to adopt a conventional applied mathematics approach for text mining, within which documents are described in a vector space model. During this model, every document is described by a vector containing the frequencies of occurrences of words that are outlined as delimited alphabetic strings, whose range of characters is between four and twenty five. We also used a dimensionality reduction technique called Term Variance (TV) that may increase both the effectiveness and potency of clustering algorithms. TV selects variety of attributes (in our case a hundred words) that have the best variances over the documents. So as to calculate distances between documents, two measures are used, namely: cosine-based distance and Levenshtein-based distance. In order to estimate the amount of clusters, a wide used approach consists of obtaining a group of data partitions with totally different numbers of clusters and so choosing that specific partition that gives the most effective result according to a particular quality criterion (e.g., a relative validity index). Such a collection of partitions might result directly from a hierarchical clustering dendrogram or, or else, from multiple runs of a partition algorithm (e.g., K-means) ranging from totally different numbers and initial positions of the cluster prototypes. Query process is simply like search engine in the web, the results are the based upon the page ranking of the documents, similarly in our system results (i.e. documents) are show based on term frequency.

V. IMPLEMENTATION DETAILS

A. Preprocessing Steps:

Preprocessing of text documents is important to clean data and provide algorithms solely the desired data. It takes input as a text document and output a collection of tokens (which may be single terms or n-grams).

The preprocessing techniques used in our system are described below:

1. **Tokenization:** It takes text as input and outputs the number of tokens.
2. **Removal of Stop Words:** we tend to maintain a stop word table having all potential stop words. We scan our documents to seek out such stop words and take away it.
3. **Stemming**: Once stop word removal, we performed stemming of words. We maintained indexed stems. For first index position we kept the original stem, and then we tend to scan the document to form the stems. For example: bail / bailed / bailing. So, if we found any word like bailed or bailing then we replace these with bail
4. **Weighted matrix construction:** It involves the development of weighted matrix primarily based upon the frequency of occurrence of words.

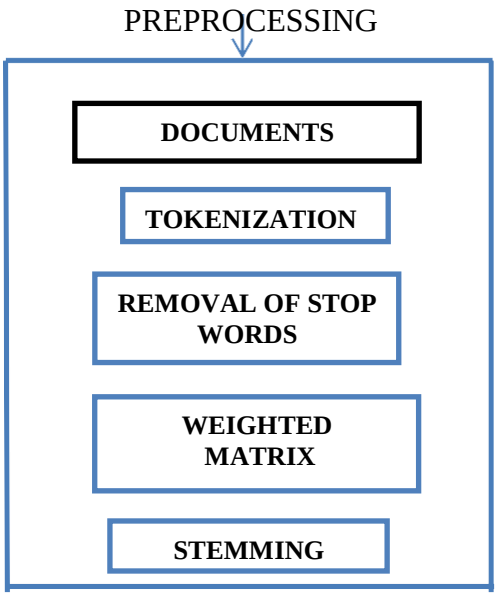

Fig 2, Preprocessing of Unstructured Documents

B. Clustering Algorithms:

K-means and Incremental mining are the famous Algorithms in the machine learning and data mining fields, and therefore they have been used in our study.

K-means Algorithm:

K-means starts with selection of K randomly chosen objects as initial clusters centers, named as seeds. The cluster centers are moved around in space in order to Minimize the RSS. These two steps are repeated iteratively until a stopping criterion is met.

- Reassignment of objects is done to the cluster with the closest centroids.
- Each centroid is recomputed based on the current members of its cluster.

The termination conditions as stopping criterion are:

- The numbers of iterations are equal to a pre-decided value for number of iterations to be completed.
- The centroids μi are not toggling between iterations.

- Termination of algorithm when the RSS value falls below a pre-established threshold.

Algorithm for K-Means

1. **Procedure** KMEANS(X,K)
2. {s1, s2, · · · , sk} Select Random Seeds(K,X)
3. **for** i ←1,K **do**
4. μ(Ci) ← si
5. **end for**
6. **repeat**
7. mink~xn−~μ(Ck)k Ck = Ck [{~xn}
8. **for all** Ck **do**
9. μ(Ck) = **1**
10. **end for**
11. **until** stopping criterion is met
12. **end procedure**

Cosine Similarity Measure: Measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them.
The cosine of 0° is 1 and for any other angles it is less than 1. It is a judgment of orientation and not of magnitude:

- Hence, the two vectors with the same orientation have a cosine similarity of 1.
- The two vectors at 90° and two vectors diametrically have a similarity of 0.
- Opposed have a similarity of - 1independent of their magnitude.

Cosine similarity is used in positive space particularly, where the outcome is neatly bounded in [0, 1].

VI. ADVANTAGE

• *Finding Similar Documents*: This feature is commonly used once the user has noticed one ‖good‖ document during a search result and desires more-like-this. The fascinating property here is that clustering is able to find documents that are conceptually alike in contrast to search-based approaches that are solely able to discover whether or not the documents share several of similar words.
• *Organizing Large Document Collections*: Document retrieval focuses on finding documents relevant to a specific query, however it fails to resolve the problem of creating sense of an oversized variety of unsorted documents. The challenge here is to arrange these documents in taxonomy just like the one humans would produce given enough time and use it as a browsing interface to the original collection of documents.
• *Search Optimization*: Clustering helps plenty in rising the standard and potency of search engines because the user query may be initial compared to the clusters rather than comparing it directly to the documents and therefore the search results may also be organized easily.

VII. CONCLUSION

The paper presents the analysis of documenting clustering techniques like partitioned clustering and hierarchical clustering. K-means presents the partitioned clustering and Hierarchical clustering makes an attempt to form a hierarchical decomposition of the given document collection therefore achieving a hierarchical structure. Similarity measures can be used to outline the performance of clusters. This approach may be very helpful for organizing vast unstructured information into structured information and enhance the method of document examination.

VIII. REFERENCES

[1] L. Filipe da Cruz Nassif, –Document clustering for forensic Analysis‖ IEEE Transaction on Information forensics and Security, 2015.
[2] B.S Everitt, s. Landau, and M. Leese, –cluster Analysis,‖ London U.K 2005.
[3] A. K. Jain and R.C Dubes, –Algorithm for clustering data‖.
[4] C. M Bishop, Pattern Recognition and Machine Learning. New York 2006.
[5] E. R. Hruschka, R. J. G. B. Campello –an approach for extraction of news content using similarity‖
[6] N. L. Beebe and J. G. Clark –Digital forensic text string searching, Elsevier 2012.
[7] Hadjidj, M. Debbabi, H. Lounis, –Towards an integrated e-mail forensic analysis frame-work‖ Elsevier, vol. 5, 2009.
[8] Jian Ma, Wei Xu, Yonghong Sun, Efraim Turban, Shouyang Wang, and Ou Liu, — An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection,‖ IEEE Trans. Syst., Man, Cybernetics., A, Syst., Humans, vol. 42, no. 3, pp. 784–790, May. 2012.
[9]. S. Decherchi, S. Tacconi, J. Redi –text clustering for digital forensic analysis‖ Intell. Security Inf. Syst. 2009
[10]. –Document Clustering for computer Inspection‖ IJETR JAN 2015.
[11] Punitha, S. C., and M. Punithavalli, –Performance Evaluation of Semantic based and Ontology Based Text Document analysis‖, Elsevier 2012.
[12] Qiujun, L. 2010, –Extraction of News Content for Text Mining Based on Edit Distance", Journal of Computational Information Systems, 2010, pp.3761-3777.