ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

Information Retrieval System through Advance Data Mining Using Clustering Techniques

K.V. Sumathi [1], R.Mohanapriya

Department of Computer Science, Selvamm Arts and Science College (Autonomous), Namakkal, India[1]

M. Phil Scholar, Department of Computer Science, Selvamm Arts and Science College (Autonomous), Namakkal, India[2]

ABSTRACT: Clustering is a useful data mining tool to handle information retrieval system can be clustered using any of the clustering algorithm such as K-means, ROCK etc. The IR systems help to retrieve necessary information from massive databases over the internet.In the emerging new wave of applications where people are the ultimate target of text clustering methods, cluster labels are intended to be read and comprehended by humans. The primary objective of a clustering method should be to focus on providing good, descriptive cluster labels in addition to optimizing traditional clustering quality indicators such as document-to-group assignment. In yet other words: in document browsing, text clustering serves the main purpose of describing and summarizing a larger set of documents; the particular document assignment is of lesser importance.

Descriptive clustering is a problem of discovering diverse groups of semantically related documents described with meaningful, comprehensible and compact text labels. In this article, to implement a new type of clustering methods for information retrieval which focuses on revealing the structure of document collections, summarizing their content and presenting this content to a human user in a compact way. In our implementation of Description Comes First (DCF) are in two clustering algorithms. First one is applicable to search results clustering and in second one is Descriptive k-Means which is applicable to collections of several thousand short and medium documents. Experimental results show the performance and scalability are more efficient than existing works.

I. INTRODUCTION

Information Retrieval

Information Retrieval (IR) is the area of study concerned with searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web. In response to various challenges of providing information access, the field of Information Retrieval evolved to give principled approaches to searching various forms of content. The field began with scientific publications and library records, but soon spread to other forms of content, particularly those of information professionals, such as journalists, lawyers, and doctors. Much of the scientific research on Information Retrieval has occurred in these contexts, and much of the continued practice of Information Retrieval deals with providing access to unstructured information in various corporate and governmental domains. Many universities and public libraries use IR systems to provide access to books, journals and other documents. Web search engines are the most visible IR applications.

Data mining is the investigation and study of huge data sets; in regulate to discern significant patternand rules. The intention of data mining is considered for, and effort preeminent with big data sets. It ishaving assorted statistics of technique which have several possess qualifications, but in this paper, we will concentrate on clustering techniques and its methods. Thus, the users need an optimized and summarized outcome as their searching result. The existing technologies of the search engine or any other searching technique was fetched the conclusions in dissimilar manner similarly adapted web search. In this planned architecture the description technique of summarizing the dataset to obtain the theme condensed data as the user's result.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

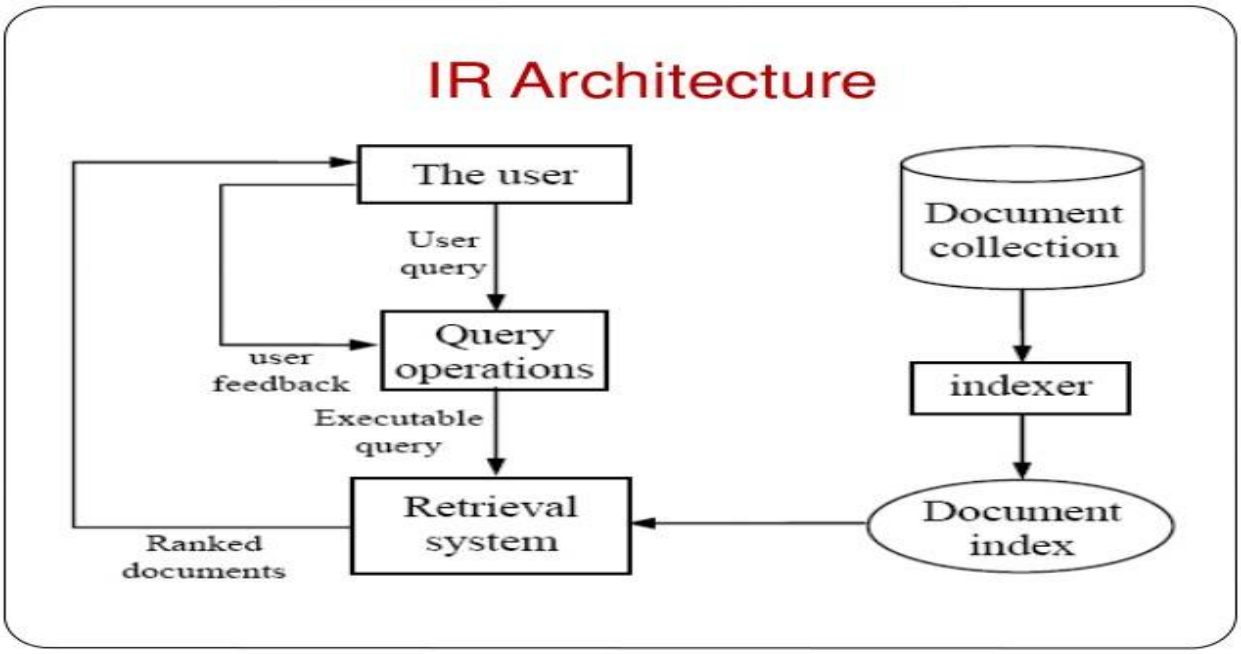

Figure 1.1 IR Architecture

Nowadays, a large quantity of data is being accumulated. Usually there is a huge gap from the stored data to the knowledge that could be construed from the data. This transition won't occur automatically, that's where Data Mining comes into picture. In Exploratory Data Analysis, some initial knowledge is known about the data, but Data Mining could help in a more in-depth knowledge about the data. Seeking knowledge from massive data is one of the most desired attributes of Data Mining. Manual data analysis has been around for some time now, but it creates a bottleneck for large data analysis. Fast developing computer science and engineering techniques and methodology generates new demands to mine complex types of data. A number of Data Mining techniques (such as association, clustering, classification) are developed to mine this vast amount of data.

Data Mining – Concepts and Techniques

Database technology has evolved from primitive file processing to the development of database management systems with query and transaction processing. Due to the explosive growth in data collected from applications including business and management, government administration, scientific and engineering, and environmental control, there is an increasing demand for efficient and effective data analysis and data understanding tools. Data warehouse systems provide some data analysis capabilities which include data cleaning, data integration and OLAP (On-Line Analytical Processing). These analysis techniques provide functionalities such as summarization, consolidation and aggregation, as well as the ability to view information at different angles. Although OLAP tools support multidimensional analysis and decision making, additional data analysis tools are required for in-depth analysis, such as data classification, clustering, and the characterization of data changes over time.

Definition 1 :

Data Mining, also referred as knowledge discovery in databases, is a process of nontrivial extraction of implicit, previously unknown and potentially useful information (such as knowledge rules, constraints, regularities) from data in databases.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

Definition2 :

"Data Mining is the process of sorting through large amounts of data and picking out relevant information. It is usually used by business intelligence organizations, and financial analysts, but is increasingly being used in the sciences to extract information from the enormous data sets generated by modern experimental and observational methods".

Architecture of Data Mining

Broadly Data Mining architecture has three layers- Database Layer with sub-layers to prepare and store data and metadata, Data Mining Application Layer which uses the algorithms to process the data and store the results in the database, Front-End Layer to facilitate the parameter settings for Data Mining Application and visualization of the results in interpretable form.

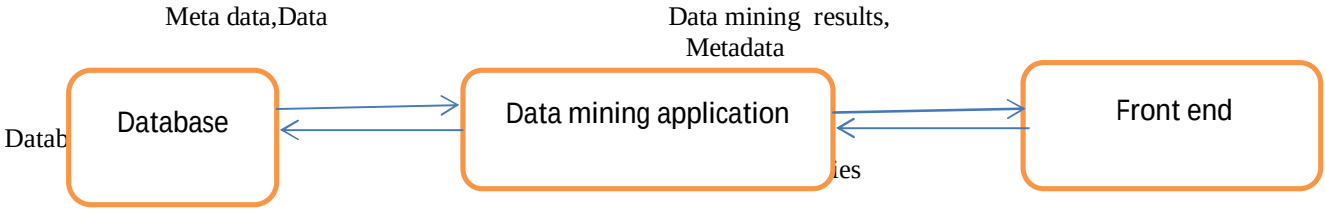

Figure 1.2 Generic 3-layer architecture for Data Mining

Nature of Data

Different data stores on which mining can be performed. In principle, Data Mining should be applicable to any kind of information repository. This includes relational databases, data warehouses, transactional databases, object-oriented and object-relational databases, spatial databases, time-series databases, text databases, multimedia databases and the World-Wide Web. The challenges and techniques of mining may differ for each of the repository systems. A brief introduction to each of the major data repository systems listed above is given below.

Relational databases

A relational database is a collection of tables. Each tuple in relational database has a unique name, consists of a set of attributes (columns or fields) and usually stores a large number of tuples (records or rows). An object in a relational table is represented by a tuple and is identified by a unique key (Primary key) and is described by a set of attribute values.Data Mining applied to relational databases allows searching for trends or data patterns. For example, Data Mining systems may detect deviations, such as items whose sales are far from those expected in comparison with the previous year. It may also analyze customer data to predict the credit risk of new customers based on their income, age and previous credit information. Relational databases are one of the most popularly available and rich information repositories for Data Mining.

Data warehouses

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site. Data warehouses are constructed via a process of data cleansing, data transformation, data integration, data loading, and periodic data refreshing. Although data warehouse tools help support data analysis, additional tools for Data Mining are required to allow more in depth and automated analysis.

Transactional databases

In general, a transactional database consists of a file where each record represents a transaction. A transaction typically includes a unique transaction identity number (trans ID), and a list of the items making up the transaction (such as items purchased in a store). The transactional database may have additional tables associated with it, which contain other information regarding the transaction, such as sales transactional database contain information about sales and keep the record of transaction date, the customer ID number, sales person ID number, and the name of branch at which the sale occurred, and so on. A regular data retrieval system fails to answer queries like ―Which items sold well

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

together?" However, Data Mining systems for transactional data can identify the relationship between different transactions, such as it can detect the set of items frequently sold together.

Review of literature

S.AnithaElavarasiet.el.publish a survey on clustering algorithms and similarity measures for categorical data and compare almost ten different clustering algorithms. Authors mainly compare accuracy and error rate for four clustering algorithms which are K-Mean, Fuzzy K-Modes, ROCK and Squeezer. Author mainly compare two algorithms ROCK and squeezer. Author take two data sets namely mushroom and vote. Squeezer works better for mushroom data set and ROCK work better for vote data set.

ManjotKauret.el.analyze web document clustering approaches using K-Means algorithm. The performance og K-Mean algorithm depends upon choice of initial clustering centers which are chosen randomly most of the time. Authors propose a technique how to find better initial cancroids and to provide an efficient way of assigning initial data points to clusters that will reduce the time complexity. Author concluded that the proposed algorithm provide better results for various data sets. But the value of k; the number of clusters will be given as input value to K-Mean which is a problem area for using K-Mean algorithm

Muhammad Adeelet.el.analyze and compare clustering techniques such as K-Mean, Agglomerative Hierarchical Clustering and Kohonenself organizing maps in clustering document to facilitate information retrieval systems. Author compare the three algorithms on the basis of many parameters. Author concluded that K-Mean K-mean is the fastest clustering technique and self organizing map is the slowest among all three techniques

II. OVERVIEW OF SELECTED CLUSTERING ALGORITHMS

Clustering analysis is a very broad field and the number of available methods and their variations can be overwhelming. A good introduction to numerical clustering can be found in Cluster Analysis or in Cluster Classification. A more up-to-date view of clustering in the context of data mining is available in Data Mining: Concepts and Techniques.

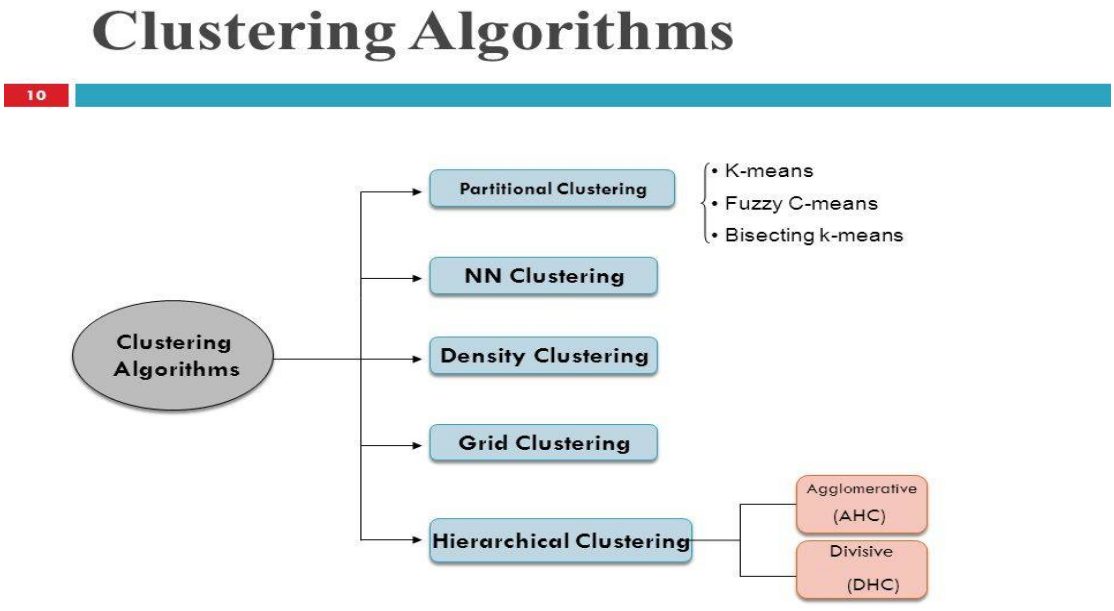

Figure 3.1 Clustering Algorithms

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

Partitioning Methods

Partitioning clustering methods divide the input data into disjoint subsets attempting to find a configuration which maximizes some optimality criterion. The most popular partitioning algorithm is the k-Means algorithm. In k-Means, we define a global objective function and iteratively move objects between partitions to optimize this function. The objective function is usually a sum of distances (or sum of squared distances) between objects and their cluster's centers and the objective is to minimize it. The representation of a cluster can be an average of its elements (its centroid) or a mean point (object closest to the centroid of a cluster).

In the latter case we call the algorithm k-Medoids. Given the number of clusters k a priori, a generic k-Means procedure is implemented in four steps:

1. Partition objects into k nonempty subsets (most often randomly),
2. Compute representation of centers for current clusters,
3. Assign each object to the closest cluster,
4. Repeat from step 2 until no more reassignments occurs.

By moving objects to their closest partition and recalculating partition's centers in each step the method eventually converges to a stable state, which is usually a local optimum. We discuss computational complexity of k-Means in later sections, for now let us just comment that the entire procedure is efficient in practice and usually converges in just a few iterations on non-degenerated data. Another thing worth mentioning is that clusters created by k-Means are spherical with respect to the distance metric; the algorithm is known to have problems with non-convex, and in general complex, shapes.

III. DATAFLOW DIAGRAM

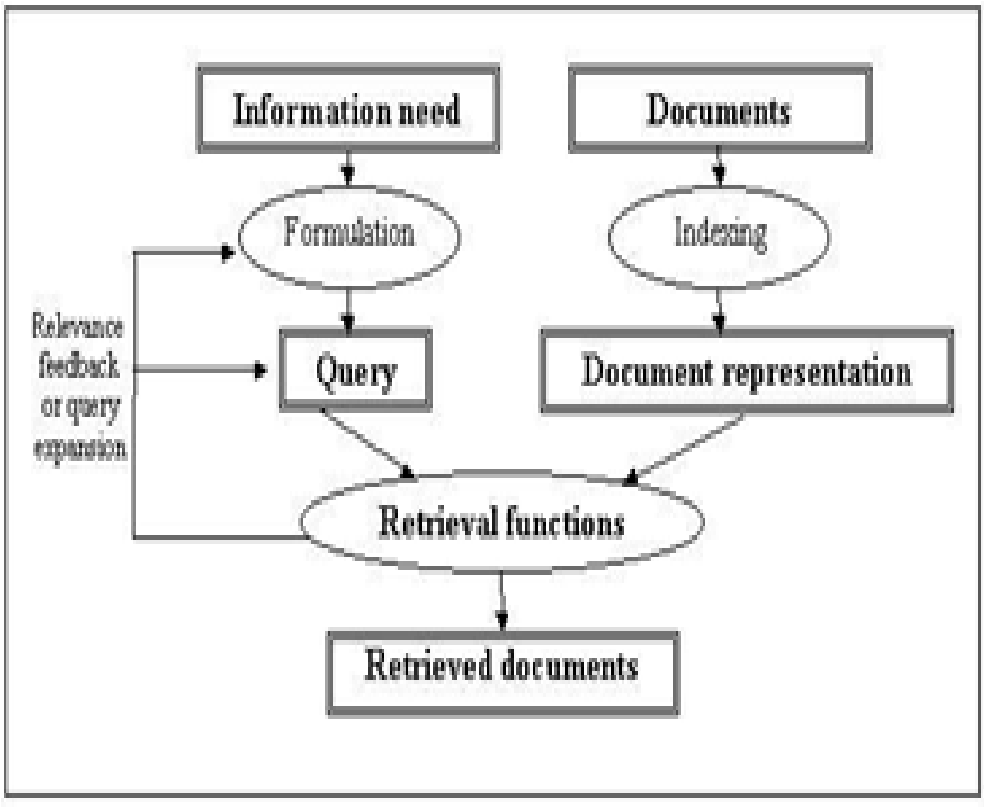

Figure 3.2 Dataflow diagram

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

Algorithm used to bootstrap k-Means

1: $D \leftarrow$ a set of input documents;
2: $k \leftarrow$ number of "topics" expected in the input;
3: $C \leftarrow$ a set of k centroid vectors c 1 . . . c k , initially empty;
4: $s = 0.1 \times ||D||$; /* Subsample ratio. */
5: Ds = random_sample (D, s); /* Select random sample of size s. */
6: $c1$ = average (Ds); /* The first centroid is the average of the sample. */
7: **for all** i = 2,3, . . . ,k **do**
8: /* Next centroid is initialized to a document most dissimilar to previous centroids. */
9: c i = argmax (Σj=1,2,3……..i-1 sim (d, c j))
d ∈D s
10: **end for**
11: C contains the initial centroids.

Initial StateProper selection of the initial state is crucial in k-Means to ensure the clusters are diverse and truly representative. We chose to initialize the algorithm by selecting most diverse documents from a subsample of the input (above).

Convergence CriterionWe used a composite convergence criterion for interrupting the computation loop in k-Means. The computation ends when any of the conditions below is true:

1. Global objective function (sum of distances from documents to their centroids,

Or

2. Fewer number of documents than min is reassigned between clusters.

The objective function in k-Means is by itself monotonic, so the algorithm will always terminate at some point. We added the second condition to trim the trailing iterations which might keep improving the global function, without affecting the centroid vectors.

Selection of the Desired Number of Clusters

We assumed that the number of dominant topics to be discovered, and at the same time the number of desired clusters for the k-Means algorithm, must be given a priori. Our decision to leave k as a parameter was partially justified by the planned experiments where k had to be given in advance to allow cross algorithm result comparisons. The second reason influencing this decision was that clustering is an internal process to Descriptive k-Means — the number of topics to be discovered can be actually set to a fixed value, if there are no sensible cluster candidates the final number of clusters should be reduced automatically anyway. Nonetheless, we are aware this is a weak point of the entire algorithm and that k could be at least estimated from the data using one of the known methods. Once k-Means converges to stable cluster centroids, they become the final result of this Phase.

IV. RESULTS EVALUATION

The goals of descriptive clustering differ slightly from those of typical clustering; we try to find coherent, properly described groups of documents, not just groups of documents. This particular aspect seemed to be completely locked for any kind of objective assessment and we knew that finding a way of performing evaluation would be a difficult issue.

Retrieval process: Issue (1):

Traditional Information Retrieval techniques become inadequate to handle large text databases containing high volume of text documents. To search relevant documents from the large document collection, a vocabulary is used

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

which map each term given in the search query to the address of the corresponding inverted file; the inverted files are then read from the disk; and are merged, taking the intersection of the sets of documents for AND, OR, NOT operations. To support retrieval process, inverted file require several additional structures such as document frequency of each lexicon in the vocabulary, term frequency of each term in the document. The principal cost of searching process are the space requirement in memory to hold inverted file entries, and the time spend to process large size inverted files maintaining record of each document of the corpus as they are potential answers.

Many terms in the query means more disk accesses into the inverted file, and more time spent to merge the obtained lists.

Issue (2):

Presently, while doing query based searching, search engines return a set of web pages containing both relevant and non relevant pages, sometimes showing non relevant pages assigned higher rank score. These search engines use one of the following approaches to organize, search and analyze information on the web. In the first approach , ranking algorithm uses term frequency to select the terms of the page, for indexing a web page (after filtering out common or meaningless words). In the second approach structure of links appearing between pages is considered to identify pages that are often referenced by other pages. Analyzing the density, direction and clustering of links, such method is capable of identifying the pages that are likely to contain valuable information.

Another approach analyzes the content of the pages linked to or from the page of interest. They analyze the similarity of the word usage at different link distance from the page of interest and demonstrate that structure of words used by the linked pages enables more efficient indexing and searching. Anchor text of a hyperlink is considered to describe its target page and so target pages can be replaced by their corresponding anchor text. But the nature of the Web search environment is such that the retrieval approaches based on single sources of evidence, suffer from weaknesses that can hurt the retrieval performance. For example, content-based Information Retrieval approach does not consider link information of the page while ranking the target page and hence affect the quality of web documents, while link-based approaches can suffer from incomplete or noisy link topology. The inadequacy of singular Web Information Retrieval approaches make a strong argument for combining multiple sources of evidence as a potentially advantageous retrieval strategy for Web Information Retrieval.

Issue (3):

A common problem of Information Retrieval is that users have to browse large number of documents containing both relevant and non relevant documents before finding relevant documents.

Clustering keeps similar documents together in a single group and hence fastens the process of Information Retrieval by retrieving the documents from the same cluster based on query vector matching to the cluster centroid.

There are many clustering algorithms available like K-means, Bisecting Kmeans, HFTC (Hierarchical Document clustering using Frequent Itemsets), Hybrid PSO+K-means method and Global K-means. But, there are many challenges in using these existing clustering techniques in the domain of text documents. Bisecting K-means produce deep hierarchy resulting in difficulty to browse if one makes an incorrect selection while navigating a hierarchy. Although HFTC produces a relatively flat hierarchy as compared to Bisecting K-means, but it is expensive in terms of calculating the global frequent itemsets to create the clusters.

Global K-means , unlike K-means, is insensitive to the choice of initial k cluster centers, thus giving global optimum solution. But it requires execution of K-means method (nk) times for document set of size n to generate k clusters showing time complexity of O(nk). In Hybrid PSO+K-means method, the PSO (Particle Swarm Optimization) module is executed for a short period to search for the optimum cluster centroid locations and then the K-means module is used for refining and generating the final optimal clustering solution. This method produces the global optimum solution like Global k-means, and also requires assuming the initial value of k.

Issue (4):

To fasten the process of document retrieval, text summarization technique is used. Ranking of documents is made based on the summary or the abstract provided by the authors of the document. But it is not always possible as not all documents come with an abstract or summary. Also when different summarization tools like Copernic, SweSum, Extractor, MSWord, Intelligent, Brevity, Pertinence text summarizer are used to summarize the document, not all the

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

topics covered within the document are reflected in its summary. Proposed Solutions The aim of this thesis is to address the above discussed issues in order to improve the accuracy of the Information Retrieval process. Our proposed approaches in this thesis contribute in the field of text Information Retrieval and provide more relevant documents (web pages) to the given searched query accurately and efficiently in less time.

- To speed up the process of text document retrieval and effective utilization of the memory space as discussed in issue (1), we propose an algorithm based on inverted index file.
- By using the range partition feature of oracle, the space requirement of memory is reduced considerably as the inverted index file is stored on secondary storage and only the required portion of the inverted index file is maintained in the main memory. Fuzzy logic is applied to retrieve the selected documents and then suffix tree clustering is used to group the similar documents.
- To handle the problem discussed in issue (2), a method is proposed for learning web structure to classify web documents and demonstrates the usefulness of considering the text content information of backward links and forward links for computing the page rank score. The similarity of the word usage at single level link distance from the target page is analyzed, which shows that content of words in the linked pages enables more efficient indexing and searching. The novel method efficiently reduces the limitations of the already existing Link Analysis algorithms like Kleinberg‘s HITS algorithm, SALSA while computing the rank score of the retrieved web pages and the results obtained by the proposed method are not biased towards in-degree or out-degree of the target page. Also the rank scores obtained showing non-zero values help to rank the web pages more accurately.
- An approach to text document clustering that overcomes the drawback of K-means and Global k-means as discussed in issue (3) is proposed which gives global optimal solution with time complexity of O(lk) to obtain k clusters from an initial set of l starting clusters.
- A new method of building the generic, extract based single text document summary of fixed length is proposed to handle the limitations of existing summarization algorithms as discussed in issue (4). Index term(s) of the document is/are identified based on keyterms of each sentence and paragraph within the document. Rank of the sentence is computed based on the number of matching terms between the document and sentence index terms. Sentences having high rank score are extracted to be included in the final summary.

V. CONCLUSION

Document clustering becomes a front-end utility for searching, but also *comprehending* information. We started this paper from the observation that clustering methods in such applications are inevitably connected with finding concise, comprehensible and transparent cluster labels a goal missing in the traditional definition of clustering in information retrieval.

We then showed that current approaches to solving the problem try to allocate a sensible description to a model of clusters expressed in mathematical concepts and not directly corresponding to the input text. This is typically very difficult and often results in poor cluster labels. Realizing the constraints and nature of the problem and having the knowledge of user expectations acquired from the Carrot framework, we collected the requirements and expectations to formulate a problem of *descriptive clustering* a document grouping task emphasizing comprehensibility of cluster labels and the relationship between cluster labels and their documents. Next, we devised a general method called *Description Comes First*, which showed how the difficult step of describing a model of clusters can be replaced with extraction of candidate labels and selection of pattern phrases labels that can function as an approximation of a dominant topics present in the collection of documents.

After exploring the literature it is concluded that web search engines and information retrieval systems provide large volume of data for given user query. Finding useful information from this large data is again a problem for the user and attracts researchers to solve this problem. Data mining technique such as clustering is a useful technique to solve this problem. The performance of different clustering techniques such as K-Mean, ROCK, Squeezer, self organizing map is analyzed in many articles. In future work can be done in applying any of the clustering technique to facilitate the information retrieval systems so that user will get the desired information easily from these systems.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 11, November 2016

REFERENCES

1) Manjotkaur and NavjotKaur, ― Web Document clustering approaches using K-Mean algorithm‖. IJARCSSE, Volume 3, Issue 5, May 2013.
2) Anoop Jain, ArunaBajpai and Manish Kumar Rohila, ― Efficient Clustering Technique for Information Retrieval in Data Mining‖. International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, Volume 2, Issue 6, June 2012.
3) S.M. Jagatheesan and V. Thiagarasu, ―Development of Fuzzy based categorical Text Clustering Algorithm for Information Retrieval‖. International Journal of Innovative Research in Computer, Vol. 2, Issue 1, January 2014.
4) S. AnithaElavarasi and J. Akilandeswari , ― SURVEY ON CLUSTERING ALGORITHM AND SIMILARITY MEASURE FOR CATEGORICAL DATA‖. ICTACT JOURNAL ON SOFT COMPUTING, JANUARY 2014, VOLUME: 04, ISSUE: 02.
5) Keole.Ranjit R and Dr.Karde.Pravin.P, ―Information Retrieval From Web Document Using Clustering Techniques‖. International Journal Of Engineering And Computer Science ISSN:2319-7242 Volume 2 Issue 3 March 2013 Page No. 759-764.
6) R.Mahalakshmi, V.LakshmiPraba, ―A Relative Study on Search Results Clustering Algorithms - K-means, Suffix Tree and LINGO‖. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-6, August 2013.
7) Muhammad Adeelet.el, ―Efficient Cluster-Based Information Retrievalfrom Mathematical Markup Documents‖. World Applied Sciences Journal 17 (5): 611-616, 2012.
8) PoonamB.Lohiya, ―A Survey On Web Search Result Clustering And Engines‖. International Journal of Science, Engineering and Technology Research (IJSETR)Volume 2, Issue 2, February 2013 ISSN: 2278 – 7798.