www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 4 Issue 1 January 2015, Page No. 10175-10179

Search Results Clustering using TF-IDF Based Apriori Approach: A Survey Paper

Hetal C. Chaudhari[1], *K. P.Wagh*[2] *and P.N.Chatur*[3]

[1]PG Scholar, Government College of Engineering,
Department of Computer Science and Engineering, Amravati. INDIA
hetalchaudhary90@gmail.com

[2]Assistant Professor, Government College of Engineering,
Department of Information Technology, Amravati. INDIA
kishorwagh2000@yahoo.com

[3]Associate professor, Government College of Engineering,
Department of Computer Science and Engineering, Amravati. INDIA
chatur.prashant@gcoea.ac.in

Abstract— ***the use of internet has increased exponentially. Search engines have become most important tool to retrieve any kind of information from the web. Users simply cast their queries on a search engine to get the desired information. More than thousands of documents are shown in search results of a query. Many times most of these web pages are not relevant to the user at all. Thus, it becomes essential for search engines to return only the relevant information to the user based on the query. Clustering makes it easier to separate out relevant results out of thousands of search results obtained for a query. Combining clustering with ranking can make it better as clustering makes groups of similar documents and applying ranking methods ranks each cluster according to their relevance with the user query. In this survey, various clustering techniques implemented before and their results are discussed. Some recent techniques discussed here are proven much more accurate than the traditional techniques.***

Keywords— web documents, search results, clustering, tf-idf, apriori algorithm

1. INTRODUCTION

In response to the user query, search engines return a list of results. The results are arranged in order of their relevance to the user query. The results obtained for a specific word can be of many different contexts according to different meanings of that word. In case of long or ambiguous queries, the resulting list may extend many vast areas on different subtopics depending on different meanings of the query. Also, there is no concrete way to know which results are relevant to the query and which are not. In this case, the results can be subdivided further and those which show similar context can be grouped together to form the clusters. Documents belonging to a group are much different than those belonging to other groups. Clustering makes a shortcut to the items that are similar in the meaning. Also, it favours systematic representation of the results. Neural networks, Association techniques, rule-based techniques, decision trees and many other techniques are used for clustering of the search results. Clustering can be characterized as the process of dividing the results in such a way that the elements belonging to a cluster are similar and those belonging to different clusters are not similar.

Clustering is performed on the search results obtained after processing a query. Here, query specific features are used for the clustering purpose. Whereas, if groups obtained after clustering are decided before getting the search results, the features which are frequent but are not relevant to the query in hand may also be taken into consideration.

Clustering subdivides the input into different regions. The input space S is divided into N no. of regions like $S_{1,}$ $S_{2...}$ S_N by calculating the similarity based on some metrics. The groups into which the input space has to be divided may not be known in advance. So, the process of clustering can also be referred as unsupervised leaning process.

2. RELATED WORK

The process of clustering groups the similar documents into a single group. The similarity among the documents is measured using some similarity metrics. Though there are many techniques available for clustering of the documents, k-means algorithm can be considered as one of the most popular clustering technique. But, the problem associated with this algorithm is that the number of clusters to be formed has to be decided first and if the choice of this number is incorrect then the results are also affected adversely.

A clustering algorithm – CBC, discovered by Patrick Pantel [1], discovers word senses from the documents. Initially, a set of clusters well occupying the input space is discovered first and the centroid of these clusters are calculated. These centroids are treated as feature vectors for the respective clusters. The words are assigned to a cluster which is most similar to the meaning of the word. Then, the overlapping features are removed so that, discovery of multiple senses of a word can be avoided. This technique of clustering faces a problem that is large amount of data is required for the training purpose to form proper clusters. The performance of this technique mainly depends on the calculation of the centroid of the cluster. Also, each time a new document is added to the input space, the centroid of the cluster changes. So, identifying proper cluster becomes difficult.

Jiyang Chen [2] proposed an unsupervised scheme for web document clustering which is based on word sense communities. In this approach, the problem of clustering the web pages is formalized as a word sense discovery problem. Suppose, a query and the search result pages are given, then according to this approach, word sense directories are detected in the extracted keyword network. Keyword extractions, finding word sense communities, community refinement, assigning documents to the labelled communities are the basic steps to implement the proposed scheme. Verbs, adjectives, adverbs, etc. are excluded while taking keywords into consideration and generally, only nouns are considered as keywords. Porter stemming algorithm is applied to perform stemming on these keywords and also the stop-words are removed. The documents are assigned to different well-formed word sense communities and thus, the clusters are formed. The necessity of the document to be clustered is measured by using the modularity score of the discovered keyword community structure. In this approach, the clusters, we get, are based on the dependency based keywords extracted for large corpus. Also, clusters are assigned labels manually.

Mansaf Alam and Kishwar Sadaf [4] proposed a technique which performs heuristic search on the query result graph. Thus, the web documents which are not desired are cut-off from the result set and clusters are formed. Also, latent semantic indexing is performed in these clusters in order to get refined, more accurate clusters having the resulting documents that are most relevant to the user query.

Li P, Wang B, Jin W. [5] proposed user-related tag expansion techniques to perform clustering of web documents. Many web documents have very few tags so using tags for clustering purpose becomes difficult. The proposed approach includes some useful tags into the original tag document. User tag data can be used as prior knowledge for this. But, adding tags in this way may change the main theme of the original document. A novel model called Folk-LDA models original and expanded tags as independent observations. This method can be applied to almost 90% tagged web documents. Also, this technique proves to be superior to the word-based methods. This indicates that tags are more useful for clustering the web documents.

Ingyu Lee, Byung -Won On [6] proposed an approach for web document clustering which uses the bisection and merge steps based on normalized cuts of the similarity graph. It deals with the skewed distributions of the cluster sizes. The traditional approaches for clustering web documents assumes that the documents are evenly distributed in the clusters or the clusters are similar in their size. The proposed approach consists of two steps. The first step is performing bisection on an affinity graph using spectral bisection. In this phase, the given graph is divided into two sub-graphs iteratively. The iteration is performed till the sub-graphs become a clique. Thus, this step creates no. of clusters that vary in sizes. In the second step of this approach, the clusters are merged together to form larger clusters. Normalized cut values are used for this purpose. The process of merging goes on until the number of merged clusters reaches to a predetermined value. The two sub-graphs are merged together to form bigger cluster only if they have the biggest normalized cut value. The sub-graphs are repeatedly merged until we get the no. of clusters equal to the desired no. of clusters.

Khaled M. Hammouda et al. [8] proposed an efficient phrase-based document indexing approach for web document clustering. The clustering of web search results is performed in two parts in this approach. The document index graph- a novel phrase based document index model is the first part of the process. Rather than relying on single term indexes only, it allows for incremental construction of the phrase based index of the document set. The emphasis is on the efficiency. If phrases are not indexed, the model could revert to the concise representation of vector space model. So, it can be said that the model is flexible. The second part of the process is an incremental document clustering algorithm. It carefully examines the pair-wise document similarity distribution inside clusters thus focusing on the perfectness of the clusters. Combination of these two parts makes a novel and robust approach for calculating similarity among the documents of the clusters which results in better results for web page clustering when compared to previously existing methods for the same. This approach analyses the web document and restructures the document in some previously determined structures. The different parts of the documents are assigned different significance levels. One of the three significance levels i.e. high, medium, low is assigned to different parts of the documents. The document is represented as a set of sentences and not as a set of words. To detect end of the sentences, a separate sentence boundary detect algorithm was developed. Weights are assigned to each sentence as per its significance. The indexing of the documents is performed using the document index graph. Phrase similarity between every pair of document is calculated and incremental clustering algorithm is applied.

3. 3. A novel design specification based K-means clustering algorithm

To improve the accuracy of clusters, Doreswamy and Hemanth K.S. [3] proposed k-means clustering approach with a novel proposed distance metric called design specification distance measure function.

K-means clustering algorithm is the most preferred algorithm for clustering even after 50 years of its discovery. In a case where distribution of the data is not known, K-means algorithm is the easiest algorithm to start with an initial number of clusters. K-means algorithm starts with deciding the number of clusters, i.e. k. Then, the centres for these clusters are decided. The next step is to attribute the closest cluster to each data point. The position of the cluster is set to the mean of all the data points belonging to that cluster. These steps are repeated until convergence. This is how the k-means algorithm generally works. The k-means algorithm works in a way that minimizes the squared error between the mean of a cluster and the data points belonging to that cluster.

Distance can be defined as the quantitative degree that enumerates the logical separation of two entities in a set based on the measurable characteristics. Distance is an amount that reflects the extent of difference/similarity between two data items. Distance is a function D with non-negative real values, defined on the Cartesian product *M* × *M* of a set R. Each point in a dataset is denoted by m_i and $n_{i.}$

The design specification distance measure function can be defined on two data points *A* and *B* as,

$$D(A,B) = \left[\sum_{i=1}^{n} |a_i - b_i|^2\right]^{\frac{p}{3}}$$

Where, n denotes the number of points in the data set. The data points are *A* = *(*a_1*,* a_2*...* a_n*) and B = (*b_1*,* b_2*...* b_n*) and p is a user defined parameter.*

To evaluate the proposed method, prototype software was designed and developed and implemented in visual C# .net. The novel design specification distance measure function in integration with the k-means algorithm was tested for clustering engineering materials database. And it was observed that the proposed algorithm with large datasets, maximizes the cluster accuracy up to 99.88% reduces the outlier accuracy to 0.019%. The performance of this method as compared to other standard methods was better.

4. A modified projected K-means clustering algorithm with effective distance measure

In high-dimensional data, the distance between a data-point and its nearest neighbour may reach its distance to the outermost data-point. In the clustering framework, the difficulty causes the distance among two data-points of the same cluster to move toward the distance among two records of various clusters. In such types of data, clusters can be present in subspaces where conventional clustering approaches cannot find them. In such cases, even if K-means algorithm is the oldest and the best method for clustering, using K-means clustering technique may not make it possible to discover subspace clusters.

So, B.Shanmugapriya et al. [9] proposed a novel modified projected K-means clustering algorithm with effective distance measure. The new algorithm generalizes the traditional K-means algorithm and the aim of the algorithm is to manage the high-dimensional data. The proposed algorithm optimizes a comprehensive objective function. The objective function of the proposed algorithm uses effective distance measure to obtain more accurate results in clustering the high dimensional data. If all the dimensions are not taken into consideration, the value of the objective function is decreased. To avoid this, the virtual dimensions are considered with the objective function. The two necessities for the data values on these virtual dimensions guarantee that the objective function attains its minimum when the clusters are found.

The objective function of the proposed algorithm is a squared error function. Suppose, $X = \{ X_{1,} X_{2,} X_{3, \dots,} X_n\}$ is the set of n data points in d-dimensional space, $C_i = (c_{i1,}\ c_{i2,} \dots, c_{id})$ is the average of the data points in the sub-set X_i and $C = \{C_{1,}C_{2, \dots,} C_k\}$, then the objective function for the traditional K-means algorithm can be obtained as follows,

$$E(c) = \sum_{i=1}^{k} \sum_{X \in x_i} \sum_{j=1}^{d} (x_i - c_{ij})^2$$

Thus, the traditional K-means algorithm ties to discover a k-partition of X that minimizes E. According to this objective function, all the dimensions take part in computation. So, it is assumed that all the clusters locate in original space.

For a subspace cluster X_i, the data points which are on an irrelevant dimension should not be considered while creating clusters. Therefore, to recognize the clusters lying in various subspaces, the concept of weight vector is introduced and a new objective function is defined.

Experiments were performed using UCI machine learning repository to evaluate the performance of the proposed algorithm. In the UCI machine learning repository, there are 211 datasets available. The performance was evaluated on the basis of two parameters, i.e. execution time and clustering accuracy and it was compared to the performance of traditional k-means approach on the same dataset. The execution time required for the modified k-means algorithm was much less for iris and wine datasets as compared to traditional approaches. Also, in case of clustering accuracy, the modified k-means approach proved itself much better than other approaches.

5. Recommendation of web pages using weighted K-means approach

To predict the user's navigational habit, a new recommendation system was proposed by R.Thiyagarajan et al. [7] the recommendation system predicts the user's browsing capacity and then recommends the web pages for a user specific to her interest. This system is based on the weighted k-means clustering algorithm.

Framework of the proposed scheme can be shown as follows.

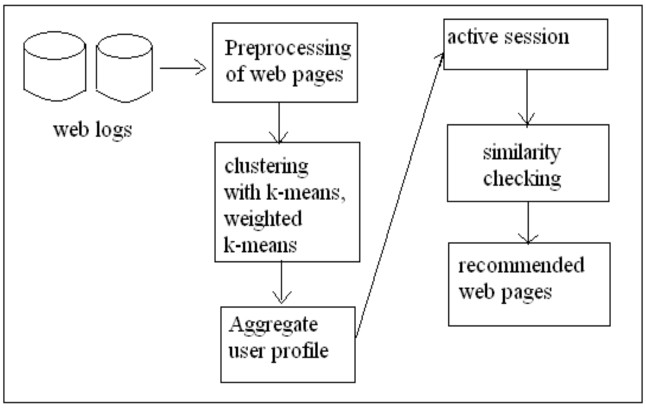

Figure 1: Framework of recommendation system

This proposed scheme performs better than traditional k-means algorithm. A recommendation list is obtained from users using weighted K-means algorithm. A framework for capturing these recommendations is generated. The recommendation list consists of the pages that are visited by the user. Also, it contains the pages visited by other users having similar usage profile. According to this algorithm, a set of weights can be associated with the data-points. These weights may have some specific meaning like they can represent the value of importance, the frequency count or any other vital information specific to the data-point. A set of objects is divided into a set of clusters that are disjoint on the basis of the fact that the numerical attributes of the objects in a set do not come from independent identical normal distribution. Weight vectors can be used to diminish the impact occurring due to irrelevant attributes in this approach. Also, the weight vectors represent the semantic information of objects. This algorithm is iterative and operates using the hill-climbing approach to get the clusters. The weights are of two types- Dynamic weights and Static weights. Dynamic weights are the weights that can be changed during the program and static weights once assigned to a data point cannot be changed during the program. The algorithm initializes with initial clusters and randomly calculates the centroids of those clusters. Static weight, ranging from 0 to 2.5 is chosen. Then, using the Euclidean distance, the distance between the centroids is calculated. The centroids are updated. The process is repeated until the new centroids are nearer to the old ones. Cosine similarity and hamming similarity measures can be used for calculating the distance between two vectors.

To evaluate the performance of this proposed scheme, experiments were carried out which used real datasets from UCI dataset repository. The performance of the proposed scheme was compared with the performance of traditional k-means algorithm. Euclidean distance measure was used to calculate the distance between different data items. And the results after the experiments showed that the hamming similarity using weighted k-means algorithm gives better results than cosine similarity measure for the binary web usage data. Weighted k-means clustering approach provides super quality recommendations for the given active user.

6. Conclusion

Web document clustering reduces the semantic mismatching between the user's actual intention behind the query and the results retuned by the search engines. In this paper, various methods for web document clustering have been reviewed.

References

[1] Patrick Pantel and D. Lin. Discovering Word Senses from Text, SIGKDD'02, July 23-26, 2002, Edmonton, Ablerta, Canada .

[2] J.Chen, O.R.Zaiane and R.Goebel. An Unsupervised Approach to Cluster Web Search Results based on Word Sense Communities, 2008 IEEE/WIC/ACM, International Conference on Web Intelligence and Intelligent Agent Technology.

[3] Doreswamy and Hemanth K.S. A Novel Design Specification (DSD) based K-mean clustering performance Evaluation on Engineering material's database, IJCA, Vol 55, No.15, Oct-2012.

[4] Mansaf Alam and Kishwar Sadf.Web search result clustering using heuristic search and latent semantic indexing, IJCA, Vol 44, No. 15, April 2012.

[5] Li p, Wang B and Jin W. Improving web document clustering through employing user-related tag expansion techniques, Journal of Computer Science and Technology 27(3):554-556 May-2012. DOI 10.1007/ S11390-012-1243-4.

[6] Ingyu Lee and B-Won. An Effective web document clustering algorithms based on bisection and merge, Artif Intell Rev (2011) 36-69-85, DOI 10.1007/S10462- 011-9203-4.

[7] R.Thiyagarajan, K.Thangavel and R.Rathipriya. IJCA, Vol-86, No. 14, Jan-2014.

[8] Khaled M and Mohamed S. Efficient phrased-based document indexing for web document clustering, IEEE Transaction on Knowledge and Data Engineering, Vol 16, No. 10, October-2004.

[9] B. Shanmugapriya and M.Punithavalli. A modified projected K-means clustering algorithm with effective distance measure, International Journal of Computer Application, Vol 44, No. 8, April-2012.

[10] Naresh Barsagade. Web usage mining and pattern discovery: A survey paper, CSE8331, Dec 2003.

H. C. Chaudhari received her B.Tech. degree in Computer Science and Technology from Usha Mittal Institute of Technology, SNDT, Mumbai, Maharashtra, India in 2011, pursuing M.Tech. Degree in Computer Science and Engineering from Government College of Engineering, Amravati, Maharashtra, India. Her areas of interest include data mining. Presently, she is engaged in web document clustering.

Prof. K. P. wagh has received his diploma in computer engineering from government polytechnic, Jalgaon, Maharashtra, India and B.E. degree in computer science and engineering from Government College of Engineering, Aurangabad. Also he has received his M.E. degree in computer science and engineering from Walchand College of Engineering, Sangli. Presently, he is working as assistant professor in Information technology department at Government College of engineering, Amravati.

Dr. P. N. Chatur has received his M.E degree in Electronics Engineering from Government College of Engineering Amravati, Maharashtra, India and Ph.D. degree from Amravari University. He has published twenty papers in international journals. His area of research includes Artificial Neural Network, Data Mining, Data Stream Mining and Cloud Computing. Currently, he is Head of Computer Science and Engineering & Electronics Engineering Department at Government College of Engineering Amravati, Maharashtra, India and is engaged with large database mining analysis and stream mining.