A Novel Intra Centroid Based Clustering for Categorical Data of Documents

Ch.M.R.Veena[1] , Dr.A.Chandra Sekhar[2]
[1]*Final M.Tech Scholor* , [2]*Associate professor*
[1,2]*Department Of Computer Science., Avanthi Institute of Engineering and Technology, A.P*

Abstract: *Keyword extraction from documents is an interesting research issue in the field of knowledge and data engineering. Every fragment or statement contains collection of keywords and may have more than one occurrence or frequency in single snippet or document. Clustering is a mechanism which groups the similar type of objects based on the similarity between the objects. In this paper we propose an efficient keyword extraction based clustering model which groups the similar type of documents based on the similarity between the documents with cosine similarity and over novel centriod computational model improves the performance of the clusters. . Our model improves the k means with elimination of random centroid selection, average pair wise distance and other parameters to generate consistent clusters our proposed model gives efficient results than traditional models.*

I. INTRODUCTION

Access to this data is molded by the accessibility of reasonable web crawlers, yet notwithstanding when these are accessible, clients regularly don't start a hunt, because their current movement does not enable them to do as such, or because they don't know that important data is accessible. We receive in this paper the viewpoint of in the nick of time recovery, which answers this inadequacy by precipitously suggesting archives that are identified with clients' present exercises.[1]

At the point when these exercises are fundamentally conversational, for example when users take an interest in a meeting, their data needs can be demonstrated as certain inquiries that are developed out of sight from the articulated words, got through constant programmed discourse acknowledgment (ASR). These understood questions are utilized to recover and suggest reports from the Web or a neighborhood store, which users can review in more detail if they discover them fascinating.

The assorted merging of recovered record records is the way toward making a short, differing and applicable rundown of prescribed documents which covers the most extreme number of themes of every discussion piece. The merging calculation rewards decent variety by diminishing the pickup of choosing documents from a rundown as the quantity of its already chose documents increments. The strategy continues in two stages. Initially, we speak to questions and the relating rundown of competitor documents from the Apache Lucene look motor utilizing theme modeling procedures, and after that we rank documents by utilizing topical closeness and compensating the scope of various records.

The ELEA Corpus comprises nearly ten hours of recorded meetings in English and French. Each meeting consists in a role play game in which participants play survivors of an airplane crash in a mountainous region. They must rank a list of 12 items with respect to their utility for surviving until they are rescued. We used from the ELEA corpus four English conversations of around fifteen minutes each, which have been manually transcribed and segmented at the speaker turn level.[2]

One of the most critical issues for a just-in-time document recommender system is to determine the appropriate timing of the recommendations, and the size of the context to use for computing them. Here, awaiting future investigations, we decided to make recommendations approximately every two minutes, at the end of an ongoing speaker turn, and consider as input the words uttered since the previous recommendation. A segment size of two minutes enables us to collect an appropriate number of words (neither too small nor too large) in order to extract keywords, model the topics, and formulate implicit queries. Based on our experience with the ACLD, it also corresponds to an acceptable frequency for receiving suggestions.

II. RELATED WORK

Even though various traditional models available in research of knowledge and data engineering, every model has its own advantages and drawbacks. Traditional clustering models cannot handle high dimensional datasets because those equality measures do not suitable for all datasets, it means if we handle with single dimensional dataset we can use Manhattan distance, for two dimensional data set we use Euclidean distance and centroid should be selected randomly, these are the major drawbacks and consistent clusters generation depends on centroid selection, so we cannot depends on simple random selection of centroids and dimensional specific clustering models

Clusters optimality completely relay on selection of centroids in traditional models

Do not support with all set of dimensional datasets.

Dimensional specific equality measures required.

Experts in the literature looking field are idealistic about the future use of intense electronic gadgets in getting more tasteful outcomes. An effective mechanical arrangement is far-fetched, in any case, assuming such present day gadgets are to be seen just as operators for quickening frameworks up to this time fitted to human abilities. A definitive advantages of automation will be acknowledged just if the qualities of machines are better comprehended and frameworks are produced which misuse these qualities without bounds. Instead of subtilize the shrewd classificatory plans now being used, new frameworks would supplant them in huge part by mechanical schedules in light of rather rudimentary thinking.[4]

Language difficulties, too, will have to be met. The problems stemming from the mere volumes of literature to be searched are being continually aggravated by the increasing accession of foreign-language documents that rate consideration on an equal level with domestic material. To be of real value, future automatic systems will have to provide a workable means of overcoming the language barrier.

The presentation of time as an extra factor will change extents impressively. In the event that, for example, any sort of data must be situated in a matter of minutes, the conceivable greatest of gifted exertion will have to be spent at the info period of the framework and in an approach degree on each passage into the framework. Assuming, be that as it may, time necessities are less squeezing, input systems that require medium ability and least exertion might be picked with the goal that the talented exertion can be gathered at the yield stage on just a little division of the records of the accumulation. In the last case, the way that lone a little division of the records of an accumulation will ever be chosen should bring about a lessening of the general exertion.

Time may affect a system in another way that makes the shift of skilled effort to the output phase more desirable. Excessive editing obviously increases the likelihood of bias due to current interests, experiences, and points of view. In consequence the usefulness of the system will be reduced as emphases and interests change. It would therefore appear that the less information is classified and contracted at the input, the more it will lend itself to dynamic interpretation at the output phase.[3]

III. PROPOSED WORK

We propose an empirical model of document clustering model with a novel clustering model which generates the centroid with intra cluster objects or documents. We compute the frequency of the keyword in all documents and relative frequency with respect to all documents. Centroid can be computed randomly from the set of objects as per the specified number of centroids and computes the cosine similarities between other objects and centroids and based on the similarity we move to cluster buckets. From the second iteration onwards we compute intra cluster centroid for more efficient clusters and continue a maximum number of iterations. Generation of consistent clusters is always an interesting research issue in the field of knowledge and data engineering. We propose an empirical model of subspace based clustering for high dimensional data or documents, traditional models work on specific dimensional datasets and they are not optimal. In this paper, we are proposing an efficient empirical clustering model for efficient grouping of similar set of data objects. Our model improves the k means with elimination of random centroid selection, average pair wise distance and other parameters to generate consistent clusters. Our proposed results show more efficient results than traditional approaches.

Intra cluster centroids gives optimal clusters than traditional random selection. Categorical data similarity can be computed with cosine similarity and it should be maximum for categorical data and minimum for numerical data.initially preprocess raw data by eliminating the unnecessary features from datasets after the preprocessing of datasets, compute the file Weight of the datasets or preprocessed feature set in terms of term frequency and inverse document frequencies and computes the file relevance matrix to reduce the time complexity while clustering datasets. We are using a most widely used similarity measurement i.e. cosine similarity

$\mathrm{Cos}(d_m,d_n)= (d_m * d_n)/\mathrm{Math.sqrt}(d_m * d_n)$

Where

dm is centroid (Document weight)

dn is document weight or file Weight from the

In the following example diagram shows a simple way to retrieve similarity between documents in at individual data holders by computing cosine similarity prior clustering as follows.

	d1	d2	d3	d4	d5
d1	1.0	0.77	0.45	0.32	0.67
d2	0.48	1.0	0.9	0.47	0.55
d3	0.66	0.88	1.0	0.77	0.79
d4	0.89	0.67	0.67	1.0	0.89
d5	0.45	0.88	0.34	0.34	1.0

Fig1: Similarity Matri

In the above table $D(d_1,d_2\ldots.d_n)$ represents set of documents at data holder or player and their respective cosine similarities, it reduces the time complexity while computing the similarity between the centroids and documents while clustering.

In our approach we are enhancing K Means algorithm with recentoird computation instead of single random selection ate every iteration, the following algorithm shows the optimized k-means algorithm as follows

Algorithm :

1: Select K points as initial centroids for initial iteration

2: until Termination condition is met (user specified maximum no of iterations)

3: get_relevance(dm,dn)

Where $d_m$is the document M file Weight from relevance matrix

$d_n$is the document N file Weight from relevance matrix

4: Assign each point to its closest centroid to form K clusters

5: Recompute the centroid with intra cluster data points (i.e average of any k data points in the individual cluster).

Ex: $(P_{11}+P_{12}+\ldots.P_{1k}) / K$

All points from the same cluster

6. Compute new centorid for merged cluster

In the traditional approach of k means algorithm it randomly selects a new centroid,in our approach we are enhacing by prior construction of relevance matrix and by considering the average k random selection of document Weight for new centroid calculation .

IV. CONCLUSION

We have been concluding over current research work with efficient clustering model over set of documents. Initially we compute the frequency of the keywords and document weights which used for similarity measurements. We generate a matrix which reduces time complexity and it works like reusable component, when every we need similarity between two objects, we need not compute multiple times. Intra centroid based computation gives more efficient clusters than traditional model because it compares more than one object at a time i.e. Centroid can be computed by multiple objects or documents.

REFERENCES

[1] M. Habibi and A. Popescu-Belis, “Enforcing topic diversity in a document recommender for conversations,” in Proc. 25th Int. Conf. Comput. Linguist. (Coling), 2014, pp. 588–599.

[2] H. P. Luhn, “A statistical approach to mechanized encoding and searching of literary information,” IBM J. Res. Develop., vol. 1, no. 4, pp. 309–317, 1957.

[3] G. Salton and C. Buckley, “Term-weighting approaches in automatic text retrieval,” Inf. Process. Manage. J., vol. 24, no. 5, pp. 513–523, 1988.

[4] S. Ye, T.-S. Chua, M.-Y. Kan, and L. Qiu, “Document concept lattice for text understanding and summarization,” Inf. Process. Manage., vol. 43, no. 6, pp. 1643–1662, 2007.

[5] A. Csomai and R. Mihalcea, “Linking educational materials to encyclopedic knowledge,” in Proc. Conf. Artif. Intell. Educat.: Building Technol. Rich Learn. Contexts That Work, 2007, pp. 557–559.

[6] D. Harwath and T. J. Hazen, “Topic identification based extrinsic evaluation of summarization techniques applied to conversational speech,” in Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2012, pp. 5073–5076.

[7] A. Popescu-Belis, E. Boertjes, J. Kilgour, P. Poller, S. Castronovo, T. Wilson, A. Jaimes, and J. Carletta, “The AMIDA automatic content linking device: Just-in-time document retrieval in meetings,” in Proc. 5th Workshop Mach. Learn. Multimodal Interact. (MLMI), 2008, pp. 272–283.

[8] A. Popescu-Belis, M. Yazdani, A. Nanchen, and P. N. Garner, “A speech-based just-in-time retrieval system using semantic search,” in Proc. Annu. Conf. North Amer. Chap. ACL (HLT-NAACL), 2011, pp. 80–85.

[9] P. E. Hart and J. Graham, “Query-free information retrieval,” Int. J. Intell. Syst. Technol. Applicat., vol. 12, no. 5, pp. 32–37, 1997.

[10] B. Rhodes and T. Starner, “Remembrance Agent: A continuously running automated information retrieval system,” in Proc. 1st Int. Conf. Pract. Applicat. Intell. Agents Multi Agent Technol., London, U.K., 1996, pp. 487–495.

[11] B. J. Rhodes and P. Maes, “Just-in-time information retrieval agents,” IBM Syst. J., vol. 39, no. 3.4, pp. 685–704, 2000.

[12] B. J. Rhodes, “The wearable Remembrance Agent: A system for augmented memory,” Personal Technol., vol. 1, no. 4, pp. 218–224, 1997.