On Time Document Retrieval using Speech Conversation and Diverse Keyword Clustering During Presentations

Shruti Bhavsar, Sanjana Khairnar, Pauravi Nagarkar, Sonali Raina, Amol Dumbare

Abstract: ***In this paper we present the idea of extracting keywords from discussions, with the point of using these words to recuperate, for each small piece of conversation and generating reports to individuals. Regardless, even a smaller piece contains a blend of words, which can be effortlessly interrelated to a couple of subjects; additionally, using a customized talk affirmation (ASR) system presents slips among them. Thus it is hard to sum up effectively the data needs of the conversation individuals. We initially propose a count to kill significant words from the yield of an ASR system which makes usage of topic showing strategies and of a sub particular prize limit which supports varying characteristics in the word set, to organize the potential contrasting characteristics of subjects and diminish ASR disturbance. By then, we set forward a strategy to surmise different topically detached requests from this definitive word set, remembering the ultimate objective is to build the potential outcomes of making at any rate one appropriate proposition while using these inquiries to investigate the English Wikipedia. The readings depict that our pronouncement continue ahead over past procedures that watch simply word recurrence or idea commonality, and states the good response for a report recommended framework to be used as a piece of conversations.***

Keywords**: Document Recommendation, Information retrieval keyword extraction, Meeting analysis, Local database, Extraction, Keyword, Clustering***

I. INTRODUCTION

The world is data generating factory, about 2.5 quintillion bytes of information or data is generated on an average single day. The data scientist report that we generate 1.7Mb of data every milli sec. This data is available in every form to be processed. To calculate the amount data available on internet we have to sum up data of Amazon, Google, Facebook, Microsoft. From writing data on leafs in ancient to storing data on cloud, man have outdone themselves in this field. Accessing this data may not be tedious task, but finding relatable data surely is. The normal queries fired on Google is around 1 million and for each query it uses 1000 computers. The data retrieved on this web browsers are not open and free to all and are not efficient. In this paper we obtain the insight on the quick recovery, which restrict this insufficiency by putting forward records that are relevant to clients work. The clients work maybe conversational like presentations or meetings, for illustration if client is in verbal discussion over a esoteric topic and need to explain his collogues about same instead of going through traditional method of digging up information he can use speech to text conversion further broadening the spectrum by forming query of claimed words and retrieving the reports based on same from static database or from dynamic database. The aim is to pull out relatable and recognizable cluster of keywords, and form queries arranged in order of significance which provide the sample output.

Revised Manuscript Received on September 25, 2020.

Shruti Bhavsar, Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering & Research, Pune, India. E-mail: shrutsbhavsar@gmal.com

Sanjana Khairnar, Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering & Research, Pune, India. E-mail:sanjana4690@gmail.com

Pauravi Nagarkar, Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering & Research, Pune, India. E-mail:pauravinagarkar5@gmail.com

Sonali Raina, Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering & Research, Pune, India. E-mail:sraina1998@gmail.com

Prof. Amol Dumbare, Assistant Professor, Department of Computer Engineering, Pimpri Chinchwad College of Engineering & Research, Pune, India. E-mail:amol.dumbare@pccoer.in

II. MOTIVATION

Sentence Clustering is the most important task for text processing. This technique can also be used in general tasks of text mining. In the previous system, clustering algorithm was performed on document level which gave the text. It motivated us, and thus we are performing a sentence level text extraction using clustering algorithm. By clustering we would spontaneously expect at least one of the clusters corresponding or similar to the query term. The main aim of clustering algorithm is to display out the most prominent and foremost sentence from the given text document according to the main document.

III. PROPOSED MODEL

The diagram added below describes the arranged system that we propose for our system of document suggestion. The colloquial data is given as input to the system. The voice will be recorded by the system and subsequently the speech will be converted to text using Automated Speech Recognition. The data is processed and partitioned into clusters. Clusters contains variant of keywords including n number of words. The extracted data contains common words and unique words; among which common words are neglected and unique words are considered. The keywords are ranked using “term frequency and inverse document frequency” for Just in Time Retrieval. The maximum ranked keyword document is selected for recommendation.

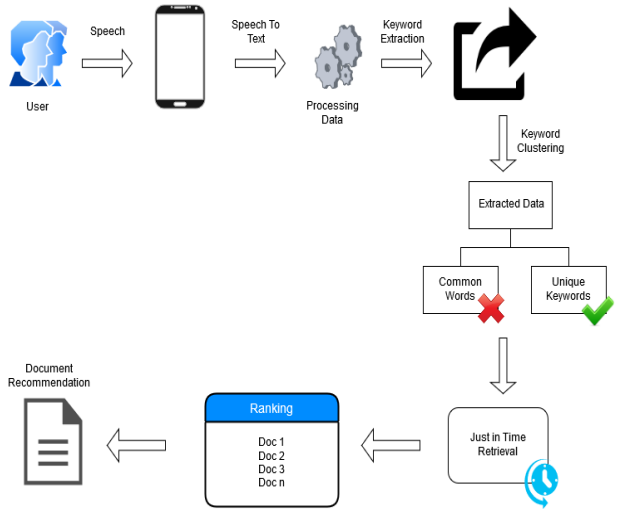

Fig.1. Architecture Diagram

IV. METHODOLOGY

A. On time retrieval system

On-time retrieval system is backbone of searching indexes all over the world. It has ability to bring on necessary and efficient changes in retrieval system. The query based retrieval system's framework persistently searches client's exercise to distinguish between data need and recovery of relatable data. This is obtained by creating a framework which consist of important words that gives the theme of overall conversation that took between clients. So we contemplate the current approach towards just-in-time retrieval framework and methods for gaining information, and studying pitfall of this system we propose our computerised content linking device(CCLD) for information retrieval during meetings.

B. Keyword extraction method

The normal orator/presenter speak around 125-130 words per minute, i.e. around 625-750 words in 5 minutes. It is found that after 5 minutes of presentation the listener can remember 50% of the content and after hour and two, he/she remembers 25% , so for normal average person it is not possible to remember every minute detail of presentation. In such scenarios the keyword extraction from recorded voice or text come handy. The keyword extraction helps in skimming through information and acquire the word that best portray the content in not more than seconds. Keyword extractions allows our clients to remember most important words from huge dataset and obtain the insight of information shared. Mechanized extraction of keywords permits you to breakdown ocean of information. Though can extract the key terms manually but this method is time consuming and inefficient. Automated keyword extraction helps to preserve the data and time. There are various techniques that can be used for keywords extractions. From statistical approach of counting words and it's frequency to machine learning approach for extracting from complex models. The paper mainly focus on structured data that contains verbal words that are noted down and further extracted from system. The various kind of factual methodology incorporate word recurrence, collocation and co event, term frequency–inverse document frequency i.e. TF-IDF and Rapid Automatic Keyword Extraction i.e. RAKE. The quibitious procedures like word recurrence, collocations and co- event consider document as "stock of words" and does not consider crucial features like structure, grammar, sequence for example synonyms cannot be detected by this method. TF-IDF measures how important the word in document It calculates term recurrence and contrast it with inverse document frequency. Multiplying both these quantities, it provides us score of word in document. Higher the score of word more it is related to theme of document. TF-IDF algorithm used in machine learning for ranking documents based on relevance of search query. So using TF-IDF for extracting relevant words based on ranking consider them as keyword is easier approach. To ameliorate recurrence based strategies linguistic and graph and lexical semantic approach have been put forward. Linguistic approach mostly is penchant towards morphological or syntactic data for example, the grammatical form of words or the relations between words in a reliance language portrayal of sentence, is utilised to figure out what catchphrases ought to be removed. The lexical semantic is subfield of linguistic. It draws meaning from text. Semantic analysis systems with support of machine learning algorithms can understand the context of natural language, detect minute details of speech and extract valuable information from complex data, achieving level of human accuracy. So in this paper we propose simple approach after studying pros and cons of variety of methods. The diverse keyword extraction from set of words captured by automatic speech recognition. Keywords are extracted with help of modelling techniques from heap of words detected and clustering those keywords to form pattern that will be accepted by queries.

C. Diverse keyword extraction

The keywords extracted from ASR have all sort of related words used while having conversation. These keywords range in all kinds. The distinct keyword selection from vast amount of available words is essential and for this there are topics modelling techniques described. Theme demonstrating is a data mining procedure that consequently examines text information to decide group words for a set of records. This is known as unsupervised machine learning since it doesn't need a predefined rundown of labels. Point demonstrating could be utilized to recognize the subjects of a lot of topics utilized in discussion by distinguishing designs and repeating words whereas the topic classification need to know that set of text before analyzing them, this set of words are provided by keyword extraction. These keywords are numbered for topic classifier to learn and make prediction later. Topic modelling refers to dividing body of document into 1) list of topics included in body of document created by ASR. 2) A few arrangements of records from the body assembled by the subjects they spread/ cover. The method used while diverse keywords extraction is latent Dirichlet Allocation i.e. LDA and latent semantic analysis i.e. LSA . The algorithm proposed for finding diverse set of keywords will extract efficient keywords from each topic that will give fair idea.

Sometimes topic created ASR noise effect belie itself as main topic of document cloaking other topics, algorithm solves this complication by choosing very few of these keywords in comparison to fallacy algorithm that ignored diversity.

D. Keyword clustering

The large collection of keywords extracted, is constructed to represent the possible information gained that might satisfy need of participant in conversation, keywords are pull-out based on theme detected in conversation. Clustering helps by collecting similar words from document which gives us precise idea of number of topics discussed in document. The reason behind clustering is to reduce ASR noise effect and increasing range of queries. Each such cluster that is related to suggested query is forwarded to document retrieval system. The algorithm that will be used for edge detection is be 'canny edge detection algorithm' which will be helped in reducing noise effects from ASR.

E. Document retrieval

As proposed, the using diverse set of extracted keywords implicit query can be prepared for every fragment of conversation. The diverse extracted keywords are clustered into groups. Using these diverse extracted keywords set, suitable document will be suggested to user based on his need and on data gathered. The accuracy is increased in these systems because of noise effects are cancelled which in turn increased efficiency and productivity.

V. FUTURE WORK

1. We optimize the time in future by using optimization technique. In our project we require 90 sec for one keyword searching, in future reducing that time by 40% using optimization technique. 2.When user enters a statement at that time stemword and stopword will be removed and only keyword will be remaining.
3.The proposed system works well for structured data , but in future we would like to apply same system for unstructured data like images and videos.

VI. CONCLUSION

The goal is to process the input (queries) resulting in rank wise document recommendation by expanding the exposure of information, while decreasing the repetition of documents. Combining these tactics in a framework should assist users to find the documents easily, thus ensuring the usability of our system. Proposed system provides complete solution for finding germane document during any verbal activity like meetings, presentations, conferences and conversations.

ACKNOWLEDGMENT

We gratefully thank the department of Computer of PCCOER, Pimpri Chinchwad, Pune for their guidance and advice in the completion of paper. We really appreciate our guide Prof. Amol Dumbare for his continuous support and valuable suggestions in our research work. Additionally, we would also like to thank all our colleagues who helped and motivated us at each step.

REFERENCES

1. A. Csomai and R. Mihalcea, "Linking educational materials to encyclopedic knowledge," in Proc. Conf. Artif. Intell.Educat.: BuildingTechnol. Rich Learn. Contexts That Work, 2007, pp. 557–559.
2. Habibi, M., Popescu-Belis, A. (2015), Keyword Extraction and Clustering for Document Recommendation in Conversations, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(4), 746–759. doi:10.1109/taslp.2015.2405482
3. Hunyadi, L. - Keyword extraction: aims and ways today and tomorrow. - In: Proceedings of the Keyword Project: Unlocking Content through Computational Linguistics. 2001.
4. Pere R. Comas and Jordi Turmo " Spoken Document Retrieval Based on Approximated Sequence Alignment"
5. Scott Deerwester ,Susan T. Dumais , "Indexing by latent semantic analysis"
6. Melville and Vikas Sindhwani, IBM T.J. Watson Research Center,Yorktown Heights, NY 105, Recommender System Prem hw}@us.ibm.com
7. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, An Introduction to Information Retrieval. Cambridge University Press, 2008
8. Khalid Al-Kofahi, Peter Jackson, Mike Dahn*, Charles Elberti, William Keenan, John Duprey.A "Document Recommendation System Blending Retrieval and Categorization Technologies".
9. Michael J. Pazzani and "Daniel Billsus, Content-based Recommendation Systems ".
10. Sangeetha. J 1, Kavitha R ," An Improved Privacy Policy Inference over the Socially Shared Images with Automated Annotation Process ", / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 3166-3169.
11. Aishwarya Singh, Bhavesh Mandalkar, Sushmita Singh , Prof. yogesh Pawar, "A Survey on User-Uploaded Images Privacy Policy Prediction Using Classification and Policy Mining",International Journal of Innovative Research in Computer and Communication Engineering .Vol. 3, Issue 9, September 2015.
12. X. Liu, X. Zhou, Z. Fu, F. Wei, and M. Zhou, "Exacting social events for tweets using a factor graph," in Proc. AAAI Conf. Artif. Intell., 2012, pp. 1692–1698.
13. A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang, "Discover breaking events with popular hashtags in twitter," in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., 2012, pp. 1794–1798.
14. A. Ritter, Mausam, O. Etzioni, and S. Clark, "Open domain event extraction from twitter," in Proc. 18th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2012, pp. 1104–1112.

AUTHORS PROFILE

Shruti Bhavsar, she is currently studying Computer Engineering from Pimpri Chinchwad College of Engineering and Research, Pune. Her area of interest is data mining and data analysis. She did a project entitled heart disease prediction system. She appears to explore greater in the field of data mining.

Sanjana Khairnar, she is currently studying Computer Engineering from Pimpri Chinchwad College of Engineering and Research, Pune. Her field of interest is Data Analysis. She has handled a project named as Car Evaluation Analysis. She looks forward to work in the field of Data Analysis.

Pauravi Nagarkar, she is currently studying Computer Engineering from Pimpri Chinchwad College of Engineering and Research, Pune. Her field of interest is in domain of data science, she will be pursuing the Masters in same. She worked on project entitled Coaching class prediction and looking forward for more exploration open doors in field of Data Science.

Sonali Raina, she is currently studying Computer Engineering from Pimpri Chinchwad College of Engineering and Research, Pune. Her field of intrigue is Business Intelligence and Data Analysis. She has taken a shot at an undertaking entitled Loan Prediction System. She looks forward to investigate more in the field of Data Analysis.

Prof. Amol Dumbare has completed B.E in Computer Engineering from Pimpri Chinchwad College of Engineering, Pune in 2012. He has also completed M.Tech from JNTU, Hyderabad, Telegana in 2015. Currently Pursuing Phd from SPPU, Pune and working as Assistant Professor for Computer Engineering Department in PCCOER, Ravet, Pune. He has Industry Experience of 1 year and Teaching experience of 7 years.