ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Keyword Extraction Using Clustering With Incremental Key Index Fast Search (IKIFS)

Chitla Roja
Pursuing M.Tech,
Department of Computer Science and Engineering, Avanthi's St.Theressa Institute of Engineering and Technology, Garividi, Andhra Pradesh, India.

Dr.R.Rajender, M.Tech, Ph.D
Director & Professor,
Department of Computer Science and Engineering, Avanthi's St.Theressa Institute of Engineering and Technology, Garividi, Vizianagaram.

ABSTRACT:
This paper addresses the problem of keyword extraction from conversations, with the goal of using these keywords to retrieve, for each short conversation fragment, a small number of potentially relevant documents, which can be recommended to participants.. We first propose an algorithm to extract keywords from the output of an JIT Retrieval system (or a manual transcript for testing), which displays whether the keyword is present in a document and how many times it is present in a document, it shows the position of the keyword. Then, we propose a method IKIFS(Incremental Key Index Fast Search) is to extract the documents based upon the filtering option , Not all the documents will be derived only the user required document will be extracted, if the user need on particular format (Pdf, text, Word etc.) this system will derived only that selected files. Then the extracted documents will be stored in DB. It is a document database that provides high performance, high availability, and easy scalability. A MDB deployment hosts a number of databases. A database holds a set of collections. A collection holds a set of documents.

INDEX TERMS:
Clustering, IKIFS (Incremental Key Index Fast Search) and UFS (User Frequency Suggestion).

OBJECTIVE:
We first propose an algorithm to extract keywords from the output of an JIT Retrieval system (or a manual transcript for testing), which displays whether the keyword is present in a document and how many times it is present in a document, it shows the position of the keyword. Then, we propose a method IKIFS (Incremental Key Index Fast Search) is to extract the documents based upon the filtering option , Not all the documents will be derived only the user required document will be extracted, if the user need on particular format (Pdf, text, Wordetc) this system will derived only that selected files. Then the extracted documents will be stored in DB. It is a document database that provides high performance, high availability, and easy scalability. A MDB deployment hosts a number of databases. A database holds a set of collections. A collection holds a set of documents. A document is a set of key-value pairs. Documents have dynamic schema. Dynamic schema means that documents in the same collection do not need to have the same set of fields or structure, and common fields in a collection's documents may hold different types of data.

2. LITERATURE SURVEY:
A Statistical Approach to Mechanized Encodingand Searching of Literary Information:
Written communication of ideas is carried out on the basis of statistical probability in that a writer chooses that level of subject specificity and that combination of words which he feels will convey the most meaning. Since this process varies among individuals and since similar ideas are therefore relayed at differentlevels of specificity and by means of different words, the problem of literature searching by machinesstill presents major difficulties. A statistical approach to this problem will be outlined and the various stepsof a system based on this approach will be described.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Steps include the statistical analysis of a collection of documents in a field of interest, the establishment of a set of "notions" and the vocabulary by which they are expressed, the compilation of a thesaurus-type dictionary and index, the automatic encoding of documents by machine with the aid of such a dictionary, the encoding of topological notations (such as branched structures), the recording of the coded information, the establishment of a searching pattern for finding pertinent information, and the programming of appropriate machines to carry out a search.

Advantages:

1. The major technical effort involved in substituting mechanical for intellectual means must, of course, be justified by the improved results obtained.

Disadvantages:

1. Language difficulties, too, will have to be met.To be of real value, future automatic systems will have to provide a workable means of overcoming the language barrier.
2.It is possible to recognize better the existence of different levels and the necessity of applying appropriately different techniques to their mechanization.

Enforcing Topic Diversity in a Document Recommender for Conversations:

This paper addresses the problem of building concise, diverse and relevant lists of documents, which can be recommended to the participants of a conversation to fulfill their information needs without distracting them. These lists are retrieved periodically by submitting multiple implicit queries derived from the pronounced words. Each query is related to one of the topics identified in the conversation fragment preceding the recommendation, and is submitted to a search engine over the English Wikipedia. We propose in this paper an algorithm for diverse merging of these lists, using a submodular reward function that rewards the topical similarity of documents to the conversation words as well as their diversity. We evaluate the proposed method through crowdsourcing. The results show the superiority of the diverse merging technique over several others which not enforce the diversity of topics.

Advantages:

1.Many studies addressed the topic diversification approach for re-ranking the retrieved results of a single query. However, these approaches are not directly applicable to multiple queries.To Overcome this, We proposed our new technique to merge lists

Disadvantages:

1.When comparing Round-robin versus SimM, the scores show the superiority of the former method when the number of conveyed topics in fragments is higher than the number of recommended documents.
2.It provides a diverse lists of documents in which documents relevant to less important topics are not displayed. However, when the number of topics is smaller than the number of recommendations, SimM provides better results.

Linking Educational Materials to Encyclopedic Knowledge:

This paper describes a system that automatically links study materials to encyclopedic knowledge, and shows how the availability of such knowledge within easy reach of the learner can improve both the quality of the knowledge acquired and the time needed to obtain such knowledge.

Advantages:

1. Here the learner can improve both the quality of the knowledge acquired and the time needed to obtain such knowledge.
2.We used to improve educational materials by automatically selecting keywords, technical terms and other key concepts and linking them to an external knowledge source,

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Disadvantages:

1. Finding the correct Wikipedia article that should be linked to a candidate keyword. Here, we face the problem of link ambiguity, meaning that a phrase can be usually linked to more than one Wikipedia page, and the correct interpretation of the phrase (and correspondingly the correct link) depends on the context where it occurs.
2.This task is in fact analogous to the problem of word sense disambiguation and our system again uses state-of-the-art techniques to address this problem.

Document concept lattice for text understanding and summarization:

We argue that the quality of a summary can be evaluated based on how many concepts in the original document(s) that can be preserved after summarization. Here, a concept refers to an abstract or concrete entity or its action often expressed by diverse terms in text. Summary generation can thus be considered as an optimization problem of selecting a set of sentences with minimal answer loss. In this paper, we propose a document concept lattice that indexes the hierarchy of local topics tied to a set of frequent concepts and the corresponding sentences containing these topics. The local topics will specify the promising sub-spaces related to the selected concepts and sentences. Based on this lattice, the summary is an optimized selection of a set of distinct and salient local topics that lead to maximal coverage of concepts with the given number of sentences.

Advantages:

1.we proposed a document concept lattice that indexes the hierarchy of local topics tied to a set of frequent concepts and the corresponding sentences containing these topics.
2.Based on this lattice, the summary is an optimized selection of a set of distinct and salient local topics that lead to maximal coverage of concepts with the given number of sentences.

Just-in-time information retrieval agents:

A just-in-time information retrieval agent (JITIR agent) is software that proactively retrieves and presents information based on a person's local context in an easily accessible yet nonintrusive manner. This paper describes three implemented JITIR agents: the Remembrance Agent, Margin Notes, and Jimminy. Theory and design lessons learned from these implementations are presented, drawing from behavioral psychology, information retrieval, and interface design. They are followed by evaluations and experimental results. The key lesson is that users of JITIR agents are not merely more efficient at retrieving information, but actually retrieve and use more information than they would with traditional search engines.

Efficient estimation of word representations in vector space:

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

3. ANALYSIS:

EXISTING SYSTEM:

In the existing system Just-in-time retrieval systems have the potential to bring a radical change in the process of query-based information retrieval.JIT Retrieval system (or a manual transcript for testing), which displays whether the keyword is present in a document and how many times it is present in a document, it highlights the position of the keyword.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

It retrieve the document based on the query, a particular part of a document are retrieved to answered that query.In this section. It reads the full document only a part relevant to a query will be displayed, we review existing just-in-time-retrieval systems and methods used by them for query formulation.

Drawbacks of Existing System:

This system will extract only a particular part of the document based on query request.

- Time consumption will be high
- Relevant document will not be displayed.
- It will cover a few key word from Each topics .
- It retrieves both relevant and irrelevant part of the documents
- It displays more uncertain words in the documents and show ambiguous questions.

PROPOSED SYSTEM:

We have proposed a IKIFS (Incremental Key Index Fast Search) to derived a documents based upon the user requirements once the keyword is given for a search this system will ask for filtering option here the user needs to filter what type of file is to be extracted like pdf or Word or video or music etc. This filtering option is used to extract the file in a particular format so that search will be fast and then time consumption will be less. the user will upload a new file which user wants to add in a storage area before uploading a file user needs to assigned a key for a particular file so that by using that key particular file will be retrieved again. This file will be uploaded in a MONGO DB. It is a cross-platform, document database that provides high performance, high availability, and easy scalability. It focuses on flexibility, power, speed, and ease of use. It works on concept of collection and document. Database is a physical container for collections. Each database gets its own set of files on the file system. A single Mongo DB server typically has multiple databases. Collection is a group of Mongo DB documents. It is the equivalent of an RDBMS table. A collection exists within a single database.

Collections do not enforce a schema. Documents within a collection can have different fields. Typically, all documents in a collection are of similar or related purpose.A document is a set of key-value pairs. Documents have dynamic schema. Dynamic schema means that documents in the same collection

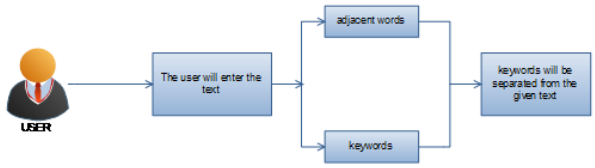
Advantages of Proposed System:

- This clustering decreases the chances errors into the queries.
- Quick response to the user.
- Content of the will be displayed

Methodologies are the process of analyzing the principles or procedure . The following are the four modules involved in the project.

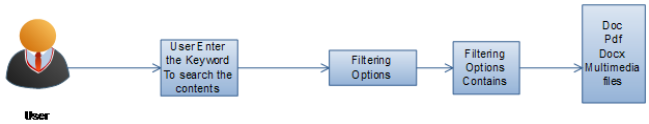
1. KEYWORD EXTRACT:

The user will enter any text inorder to extract a document,here keywords and non-keywords will be separated.Based on the keywords documents will be retrived.Then user have to select the filtering option.



2. Filtering System:

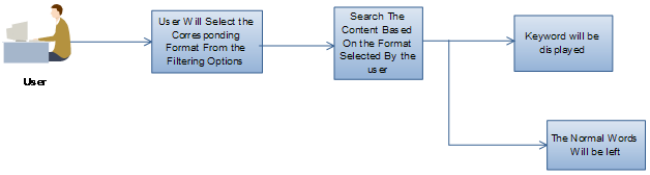
The User first enter the keyword to search the contents based on the keyword user will select at what type of file should be extracted. The filtering option contains Doc, Pdf, Docx, Music, Video files



3. Inspection system:

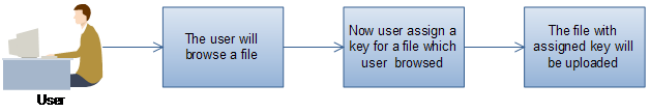
Here the user will select the corresponding format from the filtering option and search the content based on the type which is required by the user. Searches will be done based on the keywords rather than keywords, normal words will not be searched.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal



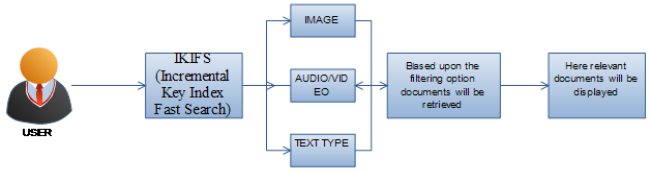
4.PERSONALISED SEARCH:

In this module the user will upload a new file which user wants to add in a storage area before uploading a file user needs to assigned a key for a particular file so that by using that key particular file will be retrieved again. This file will be uploaded in a mongo database.



SEARCH BASED ON TYPE:

Here IKIFS(Incremental Key Index Fast Search) algorithm is used to retrieve a document based on the key assigned to the document or a file.The user will select the type of the document will be retrieved i.e. text, word , pdf , image or audio/video so that relevant document will be opened.



CONCLUSION:

We addressesed the problem of keyword extraction from conversations, with the goal of using these keywords to retrieve, for each short conversation fragment, a small number of potentially relevant documents, which can be recommended to participants.. We first propose an algorithm to extract keywords from the output of an JIT Retrieval system (or a manual transcript for testing), which displays whether the keyword is present in a document and how many times it is present in a document, it shows the position of the keyword. propose a method IKIFS(Incremental Key Index Fast Search) is to extract the documents based upon the filtering option.

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," inProc. 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.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

[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.

[13] J. Budzik and K. J. Hammond, "User interactions with everyday applications as context for just-in-time information access," in Proc. 5th Int. Conf. Intell. User Interfaces (IUI'00), 2000, pp. 44–51.

[14] M. Czerwinski, S. Dumais, G. Robertson, S. Dziadosz, S. Tiernan, and M. Van Dantzich, "Visualizing implicit queries for information management and retrieval," in Proc. SIGCHI Conf. Human Factors Comput. Syst. (CHI), 1999, pp. 560–567.

[15] S. Dumais, E. Cutrell, R. Sarin, and E. Horvitz, "Implicit queries (IQ) for contextualized search," in Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2004, pp. 594–594.

[16] M. Henzinger, B.-W. Chang, B. Milch, and S. Brin, "Query-free news search," World Wide Web: Internet Web Inf. Syst., vol. 8, no. 2, pp. 101–126, 2005.

[17] D. Traum, P. Aggarwal, R. Artstein, S. Foutz, J. Gerten, A. Katsamanis, A. Leuski, D. Noren, and W. Swartout, "Ada and Grace: Direct interaction with museum visitors," in Proc. 12th Int. Conf. Intell. Virtual Agents, 2012, pp. 245–251.

[18] A. S. M. Arif, J. T. Du, and I. Lee, "Examining collaborative query reformulation: A case of travel information searching," in Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2014, pp. 875–878.

[19] A. S. M. Arif, J. T. Du, and I. Lee, "Towards a model of collaborative information retrieval in tourism," in Proc. 4th Inf. Interact. Context Symp., 2012, pp. 258–261.

[20] J. Zaino, MindMeld makes context count in search, [Online]. Available: http://semanticweb.com/mindmeld-makes-context-countsearch_ b42725 2014.

Author's Details:

Dr.R.Rajender, M.Tech, Ph.D
Director & Professor,
Department of Computer Science and Engineering,
Avanth's Institute of Engineering and Technology,
Garividi, Vizianagaram.

Chitla Roja
Pursuing M.Tech,
Department of Computer Science and Engineering,
Avanthi'S St.Theressa Institute of Engineering and Technology, Garividi, Andhra Pradesh, India.