# Survey on Keyword Extraction

**V.Haripriya[1] Mrs Ramya Devi[2]**
[1]P.G. Scholar [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]Velammal Engineering College, Chennai, India

*Abstract*—Data Mining is mainly used for storing and retrieve needed information. Reading and summarizing the contents of large entries of text into a small set of topics is difficult and time consuming for a human, so much that it becomes nearly impossible to accomplish with limited manpower as the size of the information grows. As a result, automated systems for retrieving information from large storage place, we are using the Concepts of Keyword and Keyword Extraction. It plays a vital role in the Web Based searches and normal searches of document or any important files etc in any organization. Keyword means word used in a text search, or a word in a text document that is used in an index to best describe the contents of the document. A word or phrase submitted to a search engine in an effort to locate relevant document or websites. Keyword Extraction means extracting the keyword from the implicit queries of the user. Extracting Keywords from the document helps the user to browse in fast manner, where time consumption is reduced. Keyword Extraction widely used many places like organization, college database, mail server, group discussion and internet etc.

***Key words:*** Keyword, Keyword Extraction, Retrieving Document, Less Time Consumption

## I. INTRODUCTION

Data Mining is a process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers and develop more effective marketing strategies as well as increase sales and decrease costs. Data mining depends on effective data collection and warehousing as well as computer processing. Data mining play a vital role in my project since collection of Keywords are stored in the database.

HUMANS are surrounded by lots valuable information, available as documents, databases, or multimedia resources. Access to this information is conditioned by the availability of suitable search engines, but even when these are available, users often do not initiate a search, because their current activity does not allow them to do so, or because they are not aware that relevant information is available. Inorder to provide a fast search we are using Keyword extraction technique. This is done by fragmenting the queries of the users to extract the recommended document. These fragments are stored in the Database to avoid the Keyword frequency ie to avoid the duplication of data. Keyword extraction is not that much simple because it has to relate the keyword with the user information to retrieve the recommended document. Lots of algorithm is used to extract the document by counting its frequency, word occurrence, etc. Keywords are index terms that contain most important information. Automatic keyword extraction is the task to identify a small set of words, key phrases or keywords from a document that can describe the meaning of document. Keyword extraction is considered as core technology of all automatic processing for text materials.

Now a days Keywords are extracted based on the expressions posed by human activities (like facial expressions, gaze, and their participations among groups). Thus Technology has improved a lot in extracting a keyword for searching purpose. To display the corresponding files to the user helps them to analysis the evaluation of time they spend in searching. Thus by using Keyword search even web searches becomes familiar to get the appropriate links and results for their searches. Instead of textual way of searching Keywords, advance way have implemented to search the Keyword by using voice.

## II. EXTRACTION TECHNIQUE

Various methods of locating and defining keywords have been used, both individually and in concert. Despite their differences, most methods have the same purpose and attempt to do the same thing: using some heuristic (such as distance between words, frequency of word use, or predetermined word relationships), locate and define a set of words that accurately convey themes or describe information contained in the text.

### *A. Word Frequency Analysis*

- This is used to describe the counts of keyword within the document.
- This is also called as Term Frequency-Inverse Document Frequency (TF-IDF).
- Working of this TFIDF is to weighting the term occurrence within the specific document.

### *B. Word Co-Occurrence Relationships*

- Assumption is made that a good keyword will appear frequently within the document and not within the other document.

### *C. Using a Document Corpus*

- Markov chain is used to evaluate the all the word within the corpus of the document.
- The occurrence of words frequency is determined using Probability factors(t,d)

### *D. Frequency-Based Single Document Keyword Extraction*

- Here matrix is developed to identify the word co-occurrence.
- Advantage of this method is it identifies the number of occurrence of word in the same line.

### *E. Content-Sensitive Single Document Keyword Extraction*

Key Graph is the graphical representation technique is used to cluster the word in the document. Clusters are identified by the sub graph evaluated. Candidate keywords are identified by locating nodes and edges between the separate clusters.

 

*F. Keyword Extraction Using Lexical Chains*

Lexical chains are the list of related words found in the text. Regularly used in automated text summarization. Lexical chains are easy, quickly and accurately relate to similar meanings.

*G. Key phrase Extraction Using Bayes Classifier*

To locate the Key phrases in a document, Bayes classifier are related to probability of occurrence of the key phases. Key phases are extracted from the top phrases of ranking.

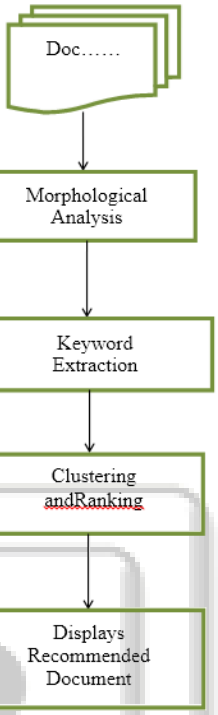


Fig. 1: Extraction Technique

## III. SURVEY DISCUSSION

Here we are going to discuss about the survey made to extract Keyword which relate the corresponding document.

*A. Just In Time Retrieval*

In this paper reviews about the Just In Time Retrieval system(JIT) which helps to extract the relevant document according to user information. Based on the context of the user Keywords are extracted to retrieve the recommended document. Example of JITR system is handling a speech to a group, sharing ideas by lecture on specific topic. 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 the 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. *Advantage*-Retrieve appropriate document to user. *Disadvantage*-Irrelevant document may be retrieved.

*B. Implicit Queries*

In this reviews about the implicit queries based on the context of the user information. The Implicit Query (IQ) prototype automatically generates queries based on user activity, and presents results in the context of ongoing work. The idea of generating background queries to retrieve task relevant information has been explored by others.[9] Thus it helps to retrieve relevant document respective to implicit queries when user is doing certain work. IQ also displays quick links to related people and topics. Users can also control presentation parameters (e.g., window size and placement, refresh delay, transparency), matching algorithms, and provide feedback about the quality of the results. Advantage-Its helps to understand the context of user queries and display the results accurately. Disadvantage-Implicit Query (IQ) prototype is not possible in all environment.

*C. Collaborative Query Reformation*

In this paper Query Reformation is discussed based on users reformulate or modify the queries when they engage in searching information particularly when the search task is complex and exploratory. This paper investigates query reformulation behavior in collaborative tourism information searching on the Web. A user study was conducted with 17 pairs of participants and each pair worked as a team collaboratively on an exploratory travel search task in two scenarios. We analyzed users' collaborative query (CQ) reformulation behavior in two dimensions:

- CQ reformulation strategies.
- The effect of individual queries and chat logs on CQ reformulation.

The findings show that individual queries and chat logs were two major sources of query terms in CQ reformulation. The statistical results demonstrate the significant effect of individual queries on CQ reformulation. It is also found that five operations were performed to reformulate the CQs, namely: addition, modification, reordering, addition and modification, and addition and reordering. These findings have implications for the design of query suggestions that could be offered to users during searches using collaborative search tools. Advantage – Thus helps the users to reform the queries according to the user turns. Disadvantage– No collaborative reformation of queries in tourism searches.

*D. Small Groups*

This illustrates about the emergent leaders among the group. The emergent leaders are the member of the team, based on their behavior and equal participants in the groups their leadership quality rise. The emergent leadership is identified by using the rule based inferences and collective classification approach. The emergent leadership of newly formed group is evolved by the nonverbal behavior, by combining speaking turns, prosodic features, visual activity, and motion. Emergent leader was perceived by his/her peers as an active and dominant person, who talks the most, has more turns and interruptions, and has a longer variation in the tone of voice and energy Advantage- Specifies the emergent leadership among the team members. Disadvantage- Judging the leaders among the team is not proper and not that much accurate.

*E. Indefinite Ranking*

This reviews about the rank list that is encountered during the searches. That discuss about priority rank given by the user while searching. A measure of the similarity between incomplete rankings should handle non-conjointness, weight high ranks more heavily than low, and be monotonic with

 

increasing depth of evaluation. The Rule Based Overlap (RBO) is based on a simple probabilistic user model. It provides monotonicity by calculating, at a given depth of evaluation, a base score that is non-decreasing with additional evaluation, and a maximum score that is non-increasing.

### *F. Learning by Reflect*

The table designed in the way to reflect, the issue of unbalanced participation during group discussions. Its displaying on its surface, a shared visualization of member participation, Reflect, is meant to encourage participants to avoid the extremes of over and under participation. We report on a user study that validates some of our hypotheses on the effect the table would have on its users. Reflect leads to more balanced collaboration, but only under certain conditions. Reflect that is designed to support collaboration between small groups. Reflect listens to the conversation taking place around it and displays information on its surface about the levels of participation of the speakers.

### *G. Temporal Templates*

The basis of the representation is a temporal template static vector-image where the vector value at each point is a function of the motion properties at the corresponding spatial location in an image sequence. Using aerobics exercises as a test domain, the representational power of a simple, two component version of the templates are explored:

- The first value is a binary value indicating the presence of motion.
- The second value is a function of the recency of motion in a sequence.

Then it is developed a recognition method matching temporal templates against stored instances of views of known actions. The method automatically performs temporal segmentation, is invariant to linear changes in speed, and runs in real-time on standard platforms. A novel representation and recognition technique for identifying movements. The approach is based upon temporal templates and their dynamic matching in time. Initial experiments in both measuring the sensitivity of the representation and in constructing real-time recognition systems have shown the effectiveness of the method.

### *H. Conversational Patterns*

The automatic discovery of group conversational behavior is a relevant problem in social computing. In this paper, we present an approach to address this problem by defining a novel group descriptor called bag of group-nonverbal-patterns (NVPs) defined on brief observations of group interaction, and by using principled probabilistic topic models to discover topics. The proposed bag of group NVPs allows fusion of individual cues and facilitates the eventual comparison of groups of varying sizes. The use of topic models helps to cluster group interactions and to quantify how different they are from each other in a formal probabilistic sense. Results of behavioral topics discovered on the Augmented Multi-Party Interaction (AMI) meeting corpus are shown to be meaningful using human annotation with multiple observers. Our method facilitates "group behavior-based" retrieval of group conversational segments without the need of any previous labeling.

### *I. Meeting Segmentation*

Multiparty meetings are a ubiquitous feature of organizations, and there are considerable economic benefits that would arise from their automatic analysis and structuring. In this paper, we are concerned with the segmentation and structuring of meetings (recorded using multiple cameras and microphones) into sequences of group meeting actions such as monologue, discussion and presentation. We outline four families of multimodal features based on speaker turns, lexical transcription, prosody, and visual motion that are extracted from the raw audio and video recordings. We relate these low-level features to more complex group behaviors using a multistream modelling framework based on multistream dynamic Bayesian networks (DBNs). This results in an effective approach to the segmentation problem, resulting in an action error rate of 12.2%, compared with 43% using an approach based on hidden Markov models. Moreover, the multistream DBN developed here leaves scope for many further improvements and extensions. The capability to incorporate some knowledge of the problem into the model structure is one of the principal features of the DBN framework, resulting in a more parsimonious model compared with simple HMMs. Moreover, the use of a multistream approach shows some advantages over merging all the feature families into a single feature vector (early integration).

## IV. PERFORMANCE EVALUATION

Performance of Keyword Extraction are measured using the all the survey made. It discuss about the efficiency of the each author, who tried to bring out their talents in extracting the keywords. Thus Keyword Extraction makes the users to evaluate the need documents easily from the surplus information.

| TITLE | PRONS | CONS |
|---|---|---|
| Just In Time Retrieval | Collects the context user information correctly | Miss understanding of user environment |
| Implicit Queries | Extracting Queries from users context | Definition for the queries is incorrectly evaluated |
| Collaborative Query Reformation | Reforming queries gives the user to get the need context information | No query reformulation behavior in collaborative tourism |
| Small Groups | Grouping team members to discuss and evaluate their participations | Accuracy of judgment by Emergent leaders |
| Indefinite Ranking | Ranking the list of user searches will provide them for fast search of documents or contents required for them | Measures of similarity are lacking using RBO |

 

| | | |
|---|---|---|
| Learning by Reflect | Users learns their conversational points and behavior while discussing certain things among them | Small groups No learning gains |
| Temporal Templates | Measuring or collecting the human non – conversional movement | Fails when more than two people in the field of view. |
| Conversational Patterns | Selection of emergent leaders among the group | Discovery of topics that suites for group patterns and the groups that best fit them |
| Meeting Segmentation | Discussion made based on the dynamic Bayesian Networks for locating the Multi stream companies | Delay in results using multi stream modelling |

Table 1: Prons & Cons

## V. CONCLUSION

As with the noun phrase keyword extraction methodology, the only requirement is that the language has a morphological analyzer and rules for finding simple noun phrases. Since nouns contain bulk of the information, noun phrases are extracted and become candidate keywords. The noun phrases are scored and clustered and then the clusters are scored. The shortest noun phrases from the highest scoring clusters are then used as the keywords.

The Position Weight algorithm automatically extracts Keywords from a single document using linguistic features. The results show that the PW algorithm has a great potential for extracting keywords, as it generates a better result than other existing approaches.

Using TF-IDF Variants, there are six different values for every word and filtering can be done by using cross domain comparison i.e. meaningless words have been removed. Furthermore, TTF(Table Term Frequency)[2] has been applied to more precise extraction of keywords.

After lots of survey made on Extraction of Recommended document, it is concluded based on the efficiency of the algorithm, there will be perfect extraction of recommended documents. So survey is going on these two algorithm IKIFS, UFS to extract the appropriate Document in accurate and Fast manner.

## REFERENCES

[1] B. J. Rhodes and P. Maes, "Just-in-time information retrieval agents,"IBM Syst. J., vol. 39, no. 3.4, pp. 685–704, 2000.
[2] 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.
[3] A. S. M. Arif, J. T. Du, and I. Lee, "Examining collaborative queryreformulation: A case of travel information searching," in Proc.37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2014, pp.875–878.
[4] D. Sanchez-Cortes, O. Aran, M. Schmid Mast, and D. Gatica-Perez,"A nonverbal behavior approach to identify emergent leaders in smallgroups," IEEE Trans. Multimedia, vol. 14, no. 3, pp. 816–832, Jun.2012
[5] W. Webber, A. Moffat, and J. Zobel, "A similarity measure for indefinite rankings," ACM Trans. Inf. Syst. (TOIS), vol. 28, no. 4, pp 20:1–20:38, 2010.
[6] K. Bachour, F. Kaplan, and P. Dillenbourg, "An interactive tablefor supporting participation balance in face-to-face collaborativelearning," IEEE Trans. Learn. Technol., vol. 3, no. 3, pp. 203–213,Jul.–Sep. 2010.
[7] A. F. Bobick and J. W. Davis, "The recognition of human movementusing temporal templates," IEEE Trans. Pattern Anal. Mach. Intell.,vol. 23, no. 3, pp. 257–267, Mar. 2001.
[8] D. Jayagopi and D. Gatica-Perez, "Mining group nonverbal conversationalpatterns using probabilistic topic models," IEEE Trans. Multimedia,vol. 12, no. 8, pp. 790–802, Dec. 2010.
[9] A. Dielmann and S. Renals, "Automatic meeting segmentation usingdynamic Bayesian networks," IEEE Trans. Multimedia, vol. 9, no. 1,pp. 25–25, Jan. 2007.

 