Context based Document Indexing and Retrieval using Big Data Analytics A Review

K.Swapnika
Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
swapnika.griet@gmail.com

K.Swanthana
Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
Swanthana.griet@gmail.com

***Abstract*—**In past few years it is observed that the internet usage is been grown wider all over the world, hence, the data generation and usage is been increased rapidly by the users, the data generated in different forms may or may not be structured. The usage of internet by individuals and organizations have been grown so, there is increasing quantity and diversity of digital data in the form of documents, became available to the end users. The Storage, Maintenance and organization of such huge data in databases is a challenging task. So, there is a great need of efficient and effective retrieval technique which focuses on improving the accuracy of document retrieval. In this paper we are going to discuss about document retrieval using context based indexing approach. Here lexical association between terms is used to separate content carrying terms and other-terms. Content carrying terms are used as they give idea about theme of the document. Indexing weight calculation is done for content carrying terms. Lexical association measure is used to calculate indexing weight of terms. The term having higher indexing weight is considered as important and sentence which contains these terms is also important. When user enters search query, the important terms are matched with the terms with higher weights in order to retrieve documents. The explicit semantic relation or frequent co-occurrence of terms is been considered in this context based indexing.

Keywords- *Context, Lexical association, Term, Weight, Semantic, Co-occurrence, Big data.*

*****

I. INTRODUCTION

Nowadays there is huge quantity and diversity of data available to the users, amount of data is present in the form of text, image, audio, video etc. Our focus is on text data or documents. Text mining deals with retrieving information from text documents. There are too much documents available in dataset and user finds difficult to get related documents he wants. So in order to ease work of user document retrieval is used. Document summarization is process in which important points from the original document are extracted. Summary of the document reduces size of the document and it gives brief idea of the document content. Overview of document can be obtained from summary of the document. There are different types of summarization like single document summarization and multi document summarization. Single document summarization is used to summarize single document and multi document summarization produces summary from multiple documents. Document retrieval is information retrieval task in which information is extracted by matching text in documents against user query. Documents related to the user query should be retrieved in acceptable time. In previous approaches there is problem of context independent document indexing. The most commonly used term weighing scheme is term frequency-inverse document frequency (TF-IDF).

Term frequency (TF): It is the frequency of term in a document. The number of times that term t occurs in a document.

Inverse document frequency (IDF): It is measure of how much information the term gives and it is given by dividing the number of documents by number of documents containing that term.

$$IDF(t) = \log(N/dft)$$

N is Total number of documents in collection, dft is number of documents with term t in it. The TF-IDF is product of TF and IDF and it is given as,

$$TF\text{-}IDF = TF*IDF$$

TF-IDF is generally used weighting factors. TF-IDF value increases proportionally as term appear in the document. The term having greater TF-IDF is considered as important in the document. In this paper effective approach is discussed for indexing the documents for providing accurate results to users query. Co-occurrence measures gives idea about how the terms are associated with the other terms in the document. Lexical association is necessary because it gives meaning and idea about theme of document. Lexical association is used to separate content carrying terms and background terms. The association between background terms is very low as compared to association between content carrying terms. The content carrying terms are assigned indexing weight according to lexical association measure. Sentences are assigned importance according to indexing weight of terms containing in it. The summary will be prepared using context based document indexing approach. Summary is used for information retrieval using TW-IDF [4] term weighting approach.

II. ALGORITHAMIC APPROACH AND SYSTEM MODEL

We are considering context of the document for summarization and graph of word approach for retrieving relevant documents. Traditional model rely on bag-of-word representation of document and scoring function TF-IDF. We are using context based document indexing approach for document summarization and graph of word and TW-IDF approach for information retrieval. Security is provided to the system by login functionality. Initially user has to register and login to the system. User has to select documents from collection. Using context based document indexing approach the summary of the document will be prepared. The generated summary of each document will be considered for the information retrieval process. Terms in summary are used for graph of word representation and the scoring function TW-IDF is used for information retrieval according to query of user. As per query of the user, relevant documents containing query terms are to be retrieved. Algorithm for the system is as shown below.

The algorithm here speaks about firstly, considering a document from document collection, and find the probability of co occurrence of terms by lexical association, Calculate indexing weight of the terms with lexical association measure and sentence score according to weight of terms, then prepare summary of document. The term weighting is done through graph of words approach and TW-IDF.

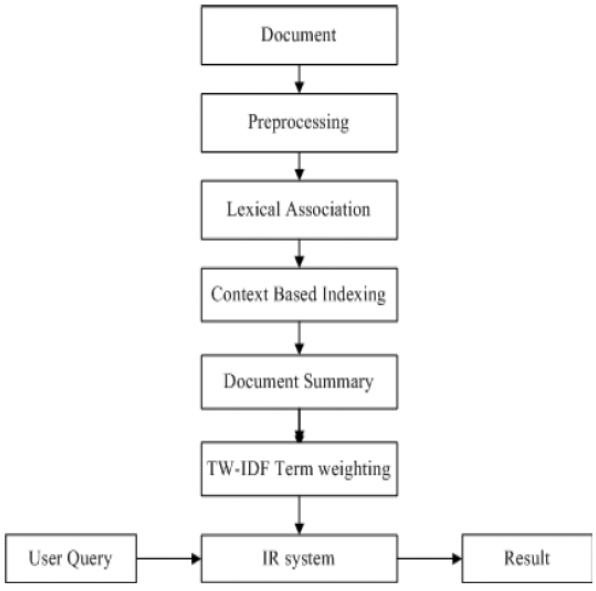

Figure 1. System Model

A. *Preprocessing*

The document which contains unnecessary words or terms such as symbols, stop words are removed or filtered. In preprocessing stage these terms from document are ignored. Preprocessing is a necessary process in order to get a condensed form of a document. Only the necessary information is provided for further stages. Preprocessing is applied on original document for summary step and then it is also applied on summarized document for graph of word approach.

B. *Lexical association*

Lexical association is a second stage, where the necessary information and meaning of document can be known. Lexical association provides us useful information and meaning of document. In lexical association background terms and the content carrying terms are separated. Content carrying terms will provide an idea about theme of the document whereas background terms give information about background knowledge. Lexical association between two content carrying terms should be more than lexical association between two background terms or between content carrying term and background term. The terms co-occurrence knowledge is used for lexical association measure. The association between topical terms i.e. content carrying terns is greater than non topical terms i.e. background terms. Thus topical terms are important which gives much information about document content. In this step pairs of consecutively written words are found from corpus and their weight is calculated considering context.

C. *Context based indexing*

With lexical association measure the topical terms in the document are given indexing weight. Here indexing. Is done through the identified topical terms, which have higher indexing weight are considered. Indexing weight of a term is calculated using context based word indexing algorithm. Indexing weight of term shows how important the term is in the document. The documents which have the important high scoring terms are retrieved as per search query. Weights of bigrams are input to Context based indexing step. The high score is assigned to the sentences having highest score. Sentence weight is calculated by summing terms other than stop words in it. By using this approach summary of document is been prepared.

D. *Term graph and TW-IDF weighting*

Here summarized document is considered for Term graph is construction, which is done by using Graph-of-word approach. Input to this step is summarized document generated using context based document indexing approach. The in-degree of the term-graph vertices is considered and the terms are given weights accordingly.. TW-IDF based retrieval model is having effective results in comparison with traditional TF-IDF approach.

III. RESULTS AND ANALYSIS

A Big Dataset which has 20,000 corpuses, we used for experiment a set of N corpus of text documents. The results of retrieval using scoring function TW-IDF will be better than traditional TF-IDF [4]. Table 1 shows comparison of Number of terms in document considered by proposed system and existing system for computation. Numbers of terms are less in proposed system as compared to original document, time required for computation is minimum.

For example, Thus performance of system improves with this approach.

Approach	**Doc 1 of corpus**	**Doc 2 of corpus**	**Doc 3 of corpus**	**…… ….**	**Doc N of corpus**	**Computation time**
No. of terms in traditional TF-IDF	229	305	260		300	Normal
No. of terms in scoring function TW-IDF	157	238	206		242	Less

Table 1: Result analysis

The system has almost 30 % less no of terms than traditional approach. This results in minimum time for calculating score of the query terms.

IV. CONCLUSION

This approach uses lexical association between terms to separate content carrying terms and background terms. The terms in the document are given indexing weight according to lexical association measure. Co-occurrence pattern between terms gives useful idea and it is used for lexical association. Summary gives condensed form of document thus minimum terms in the document which will be used for processing in next step. Here scoring function TW-IDF using term graph approach has more effectiveness than traditional TF-IDF. Here through this context based indexing approach we are going to detect the topic focusing on mining the explicit semantic relations or frequent co-occurrence relations. In future there is a need of finding out the implicit or Latent Co-occurrences of the terms in the document to reveal the important and hidden context or topic.

REFERENCES

[1] B. Pawan goyal, Laxmidhar behera, Thomas Martin Mcginnity,"A Context-based word indexing model for document summarization", IEEE Transactions on knowledge and data engineering, VoI.25,No.8,P P.1693-1705,Aug.2013,DOI: 10.1 109/TKDE.2012.1 14

[2] Jiashu Zhao, Jimmy Xiangji Huang, Ben He, ―CRTER: Using Cross Terms to Enhance Probabilistic Information Retrieval‖, SIGIR'11, 2011, pp-155-164.

[3] Jiashu Zhao, Jimmy Xiangji Huang ,―An Enhanced Context-sensitive Proximity Model for Probabilistic Information Retrieval‖, pp. 1131-1134, http://dx.doi.org/10.1145/2600428.2609527, 2014.

[4] François Rousseau, Michalis Vazirgiannis, ―Graph-of-word and TW-IDF: New Approach to Ad Hoc IR‖, pp. 59-68, http://dx.doi.org/10.1145/2505515.2505671, 2013.

[5] Osman A. S. Ibrahim , Dario Landa-Silva ―A New Weighting Scheme and Discriminative Approach for Information Retrieval in Static and Dynamic Document Collections‖, pp. 1-8, DOI: 10.1109/UKCI.2014.

[6] Veningston. K, Shanmugalakshmi. R, ―Information Retrieval by Document Re-ranking using Term Association Graph‖, http://dx.doi.org/10.1145/2660859.2660927, 2014.

[7] R. Blanco and C. Lioma, ―Random walk term weighting for information retrieval‖, SIGIR 07, pp. 829830, 2007.