International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 12 6531 - 6535

Information Retrieval Using Context Based Document Indexing and Term Graph

Mr. Mandar Donge
ME Student,
Department of Computer Engineering,
P.V.P.I.T, Bavdhan,
Savitribai Phule Pune University,
Pune, Maharashtra, India
mandardonge@gmail.com

Prof. V. S. Nandedkar
Professor,
Department of Computer Engineering,
P.V.P.I.T, Bavdhan,
Savitribai Phule Pune University,
Pune, Maharashtra, India
vaishu111@gmail.com

Abstract— Information retrieval is task of retrieving relevant information according to query of user. An idea is presented in this paper about document retrieval using context based indexing and term weighting approach. Here lexical association is used to separate content carrying terms and background 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. The summary of document is prepared. The graph of word approach is used here for information retrieval. The terms are weighted according to in-degree of vertices in document graph. When user enters search query, the important terms are matched with the terms with higher weights in order to retrieve documents. The documents which are relevant are retrieved according to weight of terms. Weight of term is determined using term graph. Term weight – Inverse document frequency scoring function is used to retrieve relevant documents. Using this approach information can be retrieved efficiently. Performance of retrieval will be improved as time required to search documents is less using proposed approach.

Keywords- *Information retieval; document indexing; lexical association; term graph.*

*****

I. INTRODUCTION

Nowadays there is huge amount of data present in the form of text, image, audio, video etc. Our focus is on text data. Text mining deals with retrieving information from text documents. Generally information retrieval is used to retrieve related information such as documents with respect to user query in short response time. 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.

$$IDF(t) = \log(N/df_t) \quad (1)$$

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

$$TF\text{-}IDF = TF*IDF \quad (2)$$

TF-IDF is generally used weighting factor in information retrieval. 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 proposed for retrieving documents related to users query. Probability of concurrence of term is found by Bernoulli model of randomness. Cooccurrence 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. In Graph of word approach the terms are assigned weights according to in-degree of vertices. The documents which are relevant are retrieved according to the query of user. Documents can be retrieved in effective way using this approach.

The rest of the paper is organized as follows. In section II, Literature survey is presented. In section III describes proposed approach, Section IV describes proposed Result and analysis, section V states conclusion related to the work.

II. LITERATURE SUREY

In "A Context based word indexing model for document summarization", Pawan goyal, Laxmidhar behera, Thomas Martin Mcginnity [1] authors have proposed context based indexing method for summarization of document. Sentence extraction based summarization is focused by the author. In this approach lexical association is used to calculate indexing weight of the terms. Bernoulli model of randomness is used for context based document indexing model. Sentence similarity matrix is calculated based on indexing weights of the terms and these are based on context.

In "CRTER: Using Cross Terms to Enhance Probabilistic Information Retrieval", Jiashu Zhao, Jimmy Xiangji Huang, Ben He [2] authors have proposed Cross term which is combination of two closely related query terms. It is term proximity approach. Cross term measures association of two terms that have textual proximity. The effect of query terms on neighboring text is approximated using kernel functions used. Cross term is overlapped effect of two terms. As distance between cross terms decreases its impact becomes lower.

In "An Enhanced Context-sensitive Proximity Model for Probabilistic Information Retrieval", Jiashu Zhao, Jimmy Xiangji Huang [3] authors have proposed probabilistic information retrieval model based on proximity. A context sensitive proximity model is proposed by integrating contextual relevance measure of term proximity to the retrieval process. Contextual relevance of term proximity represents up to what extent corresponding term pair should be related to the topic of the query. Contextual relevance of term proximity measures are integrated to propose context sensitive proximity model. Context relevance measures have been given to estimate contextual relevance of term proximity. If the term has high contextual relevance, it is assigned higher weight. The problem regarding effect of associated query term pair is focused.

In "Graph-of-word and TW-IDF: New Approach to Ad HocIR", Francois Rousseau, Michalis Vazirgiannis [4] authors have proposed graph based representation of document to model relationship between terms in the document. The document is represented as graph. The graph is unweighted and directed. In graph of word vertices are unique terms and their edges represent concurrence between terms. Vertices are the unique terms in the document. Here term weight is considered for weighting and term weight is based on in-degree of the vertices. Scoring function named TW-IDF is used for scoring "Random walk term weighting for information retrieval", R. Blanco and C. Lioma [7] is related work with graph based representation in which varying sliding window size of cooccuring terms is used as a parameter to their model.

In "A New Weighting Scheme and Discriminative Approach for Information Retrieval in Static and Dynamic Document Collections", Osman A. S. Ibrahim, Dario Landa-Silva [5] authors have proposed new term weighting scheme TF-ATO which effective Information Retrieval system. Here document centroid is used as threshold for removing less significant weights. The system proposed by author shows retrieval effectiveness in static and dynamic document collection. The TF-ATO weighting scheme shows higher effectiveness compared to TF-IDF.

In "Information Retrieval by Document Re-ranking using Term Association Graph", Veningston. K, Shanmugalakshmi. R [6] authors have proposed method of re ranking documents according to the term association and similarity among documents. Term graph is used to assess association among the words. In term graph model each document in the corpus is considered as transaction and each word is an item. Frequent item-sets of terms are found for frequent occurring terms. The proposed approach uses term graph data structure. In term graph there are two proposed approaches for re-ranking on term graph.

1. In term rank based approach the nodes of the graph are assigned rank using page rank based approach.
2. In term distance based approach, term distance matrix is constructed from term graph. Term distance is use to represent association between terms.

III. PROPOSED APPROACH

In proposed approach we are considering context of the document for summarization and graph of word approach for retrieving relevant documents. Traditional Information retrieval 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.

Algorithm

Step 1: Consider a document from document collection.
Step 2: Find probability of co occurrence of terms by lexical association.
Step 3: Calculate indexing weight of the terms with lexical association measure.
Step 4: Calculate sentence score according to weight of terms.
Step 5: Prepare summary of document.
Step 6: Term weighting using Graph-of-word approach and TW-IDF.
Step 7: Retrieve the relevant documents related to query.

Mathematical model

Set S = S1, S2, S3, S4, S5, S6, S7, S8 represents set of steps.
Where,
S1: Set of documents in collection.
S2: Preprocessing steps.
S3: Represents lexical association between terms.
S4: Context based indexing.
S5: Summary of document.
S6: Term weighting using TW-IDF.
S7: User query terms.
S8: Result of user query.

We can represent architecture using state chart diagram as follows.

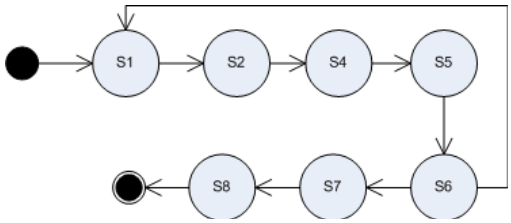

Figure 1. Mathematical Model

System model shown below gives brief idea of the proposed approach.

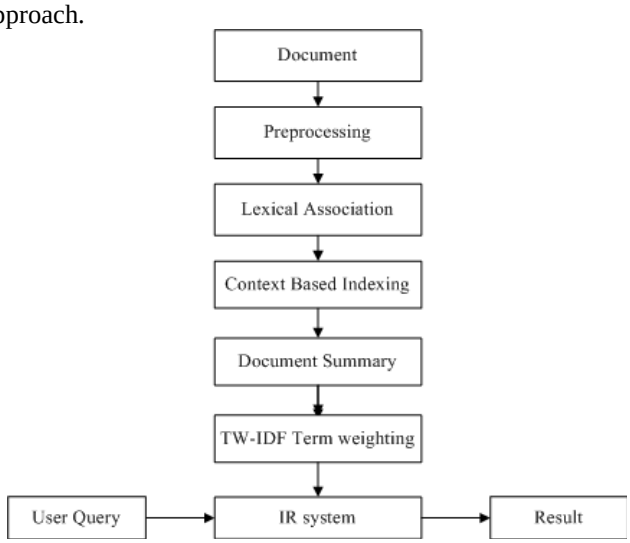

Figure 2. System model

A. *Preprocessing*

The document may contain some unnecessary information such as symbols, stop words etc. This is filtration stage. In preprocessing this unnecessary information from document is ignored. Preprocessing is necessary in order to get condensed form of the 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*

With lexical association necessary information and meaning of document can be known. Lexical association gives useful information and meaning of document. In lexical association content carrying terms and background terms are separated. Content carrying terms give more idea about theme of the document whereas background terms give information about background knowledge. Lexical association between two content carrying terms should be more the lexical association between two background terms or between content carrying term and background term. The terms cooccurrence 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 bigrams are found from all set of documents and their weight is calculated considering context of documents.

C. *Context based indexing*

With the lexical association measure the topical terms in the document are given indexing weight. The topical terms are considered for indexing. Terms which have higher indexing weight are considered. Indexing weight of 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 sentences containing higher weight terms are assigned high score. Sentence weight is calculated by summing terms other than stop words in sentence. Summary of document will be prepared using above approach.

D. *Term graph and TW-IDF weighting*

Here Term graph is prepared using Graph-of-word approach. For this summarized document is considered. Input to this step id summarized document generated using context based document indexing approach. Terms are given weights according to in-degree of the vertices in the term graph. TW-IDF scoring function is used for retrieving documents according to query of the user. TW-IDF retrieval model has effective results in comparison with traditional TF-IDF approach.

E. Information Retrieval System

Information retrieval system helps to manage and retrieve information related to the query of the user. Information retrieval system is interface for the user. User enters query for searching. Information retrieval system gets input from user in the form of query and processes the information and shows result of retrieved relevant documents. Information retrieval system calculates score of documents considering terms in query entered by user. Information retrieval system has different options for user related to operations.

IV. RESULT AND ANALYSIS

The proposed system will retrieve documents according to query of user. Summary of document will be made first and then it is considered for information retrieval. Term graph approach and scoring function is applied on summarized document instead of original document so that less number of terms will be considered for retrieval resulting in performance improvement. Final result is set of relevant documents with document id, its title and score of the document calculated using proposed approach. Dataset used for experiment has three general sample 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.

Thus performance of system improves using proposed approach.

Table 1: Result analysis

Approach	Document 1	Document 2	Document 3	Computation time
No. of terms in Existing method	229	305	260	Normal
No. of terms in Proposed method	157	238	206	Less

Fig. 3 shows the difference between the previous system and proposed method in graph format. The proposed system has almost 30 % less no of terms than existing approach. This results in minimum time for calculating score of the query terms. Thus documents can be retrieved efficiently using proposed approach of information retrieval.

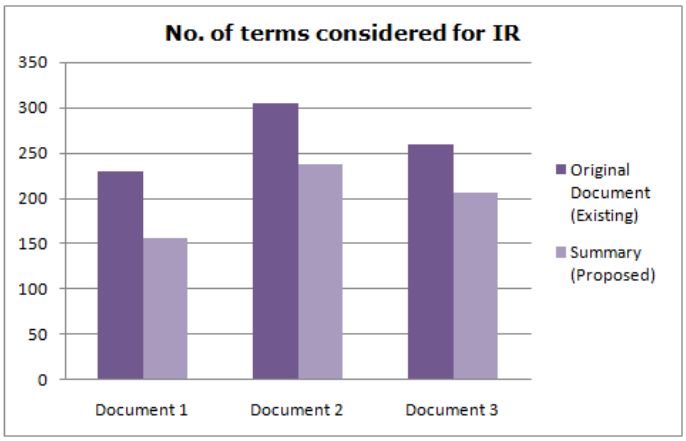

Figure 3. Result Analysis

V. CONCLUSION

In this paper we have proposed an idea for information retrieval using context based approach. The proposed 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. Cooccurrence pattern between terms gives useful idea and it is used for lexical association. Summary of the document is prepared which will be used for term weighting using graph of word approach. Summary gives condensed form of document thus minimum terms in the document which will be used for processing in next step resulting in performance improvement. Here scoring function TW-IDF using term graph approach has more effectiveness than traditional TF-IDF for information retrieval. Documents having high term weights are retrieved according to the query of the user. The documents can be retrieved effectively and in precise manner using proposed approach. Here term graph approach is applied on summarized documents instead of original document so there are less no of terms considered for information retrieval and also time required to process document is less. Currently system has been implemented as windows application. Future enhancement is to implement system as web application and improve performance of document retrieval and optimizations. Log generation can be introduced in order to monitor user's choices, search queries.

ACKNOWLEDGMENT

It is my privileges to acknowledge with deep sense of gratitude to my guide Prof. V. S. Nandedkar for her valuable suggestions and guidance throughout my course of study and timely help given to me in completion of my work. I also take this opportunity to thanks my colleague, who backed our interest by giving useful suggestions and all possible help.

REFERENCES

[1] 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.

R. Blanco and C. Lioma, "Random walk term weighting for information retrieval", SIGIR 07, pp. 829830, 2007.