International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 12 3820 – 3822

Information Retrieval using Context Based Document Indexing

Mandar Donge
[1] Department of Computer Engineering, P.V.P.I.T.
(Savitribai Phule Pune University),
Pune, Maharashtra, India
mandardonge@gmail.com

V. Nandedkar
[2] Department of Computer Engineering, P.V.P.I.T.
(Savitribai Phule Pune University),
Pune, Maharashtra, India
vaishu111@gmail.com

***Abstract*—** Information retrieval is task of retrieving relevant information according to query of user. A brief idea is presented in this paper about document retrieval using context based indexing 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. 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 importance of sentences. Using this approach information can be retrieved efficiently.

Keywords*—Information retrieval; document indexing; lexical association.*

*****

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 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 user's 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 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, related work is presented. In section III describes proposed approach, section IV states conclusion related to the work.

II. LITERATURE SURVEY

In [1] author has 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. Sentence similarity matrix is calculated based on indexing weights of the terms and these are based on context.

In [2] author proposed Cross term which is combination of two closely related query terms. It is term proximity

 ISSN: 2321-8169
3820 – 3822

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 [3] the author 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. 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 [4] the author has proposed graph based representation of document to model relationship between terms in the document. The graph is unweighted and directed. In graph of word vertices are unique terms and their edges represent concurrence between terms.
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.

In [5] author has 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 [6] the author has proposed method of re ranking documents according to the term association and similarity among documents.
The proposed approach uses term graph data structure. In term graph there are two proposed approaches

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.

III. Proposed Approach

In proposed approach we are considering context of the document for retrieving relevant documents.
As per query of the user, relevant documents containing query terms are to be retrieved. For this we are using context based indexing approach.

Algorithm
Step 1: Consider a document from dataset.
Step 2: Find probability of co occurrence 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: Retrieve the relevant documents related to query.
System model shown below gives brief idea of the proposed approach.

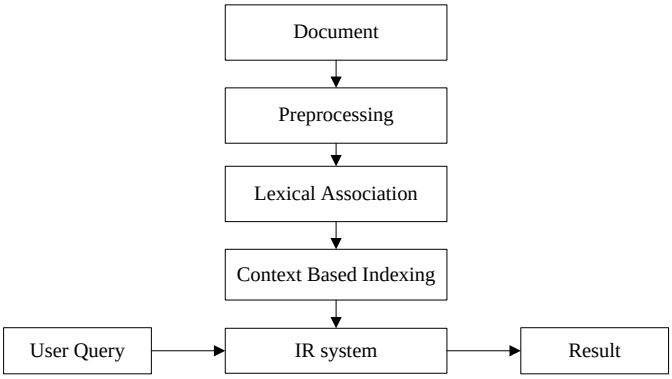

Fig. 1. System model

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.

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.

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. The sentences containing higher weight terms are assigned high score. 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.

D. 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 has different options for user related to operations.

IV. 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. Content carrying terms are important as they give much information about theme of the document. 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. Sentences are given score according to the weight of terms in sentence. Documents having high sentence score are retrieved according to the query of the user. The documents can be retrieved effectively and in precise manner using proposed approach.

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.