Retrieving Information using Context Based Indexing Approach

Ms.M.Vanishree [#1], Ms.R.Sudha [*2]

[#] *Department of Information Technology, PSNA College Of Engineering and Technology*
Dindigul,TamilNadu,India
[1]vaniishree@gmail.com

[*] *Department of Information Technology, PSNA College Of Engineering and Technology*
Dindigul,TamilNadu,India
[2]sudha@psnacet.edu.in

***Abstract*— Information retrieval is the process of retrieving the content according to the user needs. In the existing model the information retrieval is done checking the whole document. The terms related to the query are extracted. Indexing weight is applied for all the terms and finally the response is provided to the user. In the existing model they did not take the context into consideration. In this paper, we propose a context based indexing approach for retrieving the information. By using lexical association the contemporary terms and non-contemporary terms are separated. By using K-Means clustering, the contemporary terms are clustered and then the document summarization is done. Then according to the user query the information is retrieved. Then this query is matched with the summarized document. Once the keyword is matched, the particular sentence is provided as the responses to the user. Information retrieval can be done effectively by using this approach.**

***Keywords*— Indexing weight,Contemporary and Non contemporary terms**

1. Introduction

Data mining is the process of extracting the patterns from the database. Data mining act as a important tool in order to change the data into information. Extracting the information from the large database is not an easy task. In order to overcome this problem text mining is used. The amount of information are growing up now a days so the user are in needs to finds more sophisticated methods to retrieve the information effectively. Text mining is also known as text data mining and it is equivalent to text analytics. Information retrieval and lexical association are involves in the text analysis The major task involves in the text mining is to process the input text, deriving input patterns, evaluating and finally it produces the output. Information retrieval is used to support the people who all are searching for information from the database. The major problem arises in the information retrieval is that for every document the decision has to be made whether the content can be showed to the user or not. The information retrieval deals with storage, representation, organization and to access the information items. The main aim of information retrieval is to satisfy the user needs. The process of taking the sequences action to satisfy the user needs is called retrieval or searching. The information can be retrieved by the user by entering the query. Once the user enters the query, the query related information is extracted. This paper focus on retrieving the information based on context indexing approach. The contemporary term and non-contemporary terms are separated by lexical association. The contemporary terms are clustered by using k-means clustering and finally the document is summarized. Then according to the user queries the information are retrieved by using Indexing approach.

2. Related Work

Document summarization is used to produce the summary of the original document. The summarization can be done both in single and multiple documents. For summarization extraction based method is used to assign a saliency score for each sentence. Then by ranking, the sentences are selected and thus it forms the summary [2].Topic summarization is done by using content anatomy approach. Topic summarization is used to detect and track the event from the document. It organizes and summarizes the content of the temporal topic described by the set of the document. SCAN models the document as a symmetric block association matrix. Eigen vector are examined to extract events and their summaries [3].In order to extract the keyword, the keyword extraction is used. Keyword extraction is used to extract the few phrases form the text. By extracting the keyword, the summarization is done simultaneously [7]. For effective summarization the word should be classified. The word classification is done by automatic construction system and thesaurus. Thesaurus is used for word classified and hierarchy. By automatic construction method the large amount of document is classified by the semantic content [5].The word clustering and disambiguation is done based on the co-occurrences of the data. Clustering algorithm is used to merge the noun classes and verb classes. Maximum likelihood estimation is used to merge to produce the least reduction in mutual information.[6]The document summarization is done by calculating the saliency score of the sentences and by calculating the sentences similarity matrix the sentences are extracted and then the document summarization is done effectively[8].

3. System Model

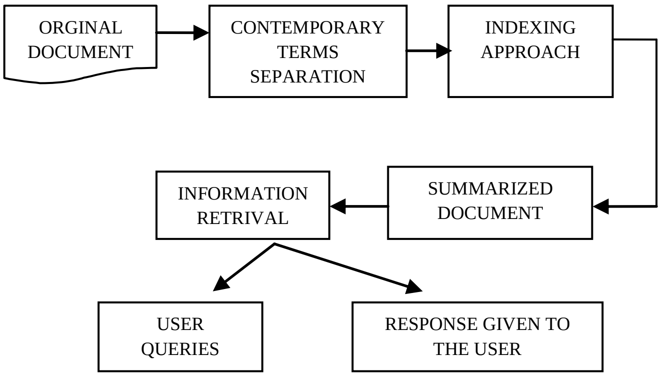


Fig 1 System overview for information retrieval

3.1 Contemporary terms separation:

Normally a document consists of lots of content and information. The terms in the document consists of contemporary terms and non-contemporary terms. The contemporary terms is used to give the important idea about the main content. The non-contemporary terms is used to give the background idea of the document. The terms are separated by using lexical association. In order to get the useful information and meaning lexical association is used.. The lexical association between the contemporary terms should be higher than the non-contemporary terms. The contemporary terms are alone extracted.

3.2 Indexing Approach:

The indexing weight is applied to each contemporary term. The term gives the highest indexing weight is chosen and finally the document summarization is done.

$$\text{Indexing weight} = tf*idf$$

3.2.1 Term frequency (tf)

In order to assign the weight to the contemporary terms term frequency is used. Term frequency can be calculated by the number of co-occurrences of the term in the document. The context-sensitive approach is used to assign weight.

3.2.2 Inverse Document frequency (idf)

Inverse document frequency is defined as the ratio between the terms in the original document by the number of terms in the summarized document and then by taking logarithm for the quotient. The term in the original document is considered as the R and the number of terms in the summarized document is considered as M.

$$idf = \log R/M \qquad (2)$$

The term which gives the highest indexing weight is taken.

3.3 Clustering:

The contemporary terms are clustered using K-Means clustering .In order to reduce the size of the document the clustering process is done. Clustering is the process of all the information. By reading the clustered document itself the user can get whole idea about the original document.

K-Means clustering:

K-Means clustering is used for clustering the items into K-clusters. The similarity of the sentences is calculated by using distances measures. The centroid in the K-Means algorithm plays an important role. The center of the cluster is called as centroid.

Algorithm:

Step 1 Choose k number of clusters.
Step2 Choose k object randomly as the initial cluster head.
Step3 Repeat the two steps,
3.1 Assign each object to the closest cluster
3.2 Compute new cluster i.e.) calculate mean point
Step 4 Stop when there is no changes in clusters head or no object change its cluster as per the stopping criteria

3.4 Document Summarization

The main aim of the summarized document is to give the overall idea about the original document. Single document summarization is used to summarize the content of the single document. The multi document summarization is used to summarize the content of the multiple documents. It is the process of extracting the main sentences from the original document. Sentences will be automatically summarized once the clustering process is finished.

3.5 Information retrieval

Information retrieval is the process of finding the information relevant to the user needs. When the user wants to retrieve the large amount of information the problem called information overload is araised. In order to reduce this problem automated information retrieval system is used. When the user enters the query the information retrieval process is done. Retrieving the information with less response time is a challenging process.

3.5.1 Queries

By using query the user can retrieve the particular information from the large database or from any document. The information retrieved should be relevant to the query. The SQL is used to retrieve and execute the appropriate keyword from the document.

3.5.2 Indexing algorithm

Indexing algorithm is used for searching the text quickly. Precision and recall is used to measure the performance of the information retrieval system. In order to evaluate precision and recall, the document collection and a query is needed. It is also used to extract the information accurately and effectively. The algorithm is as follows.

Algorithm

Step 1 The query q is entered by the user.
Step 2 Precision and recall is calculated for the query.
Step 3 Then the term related to the query is searched form the summarized document.
Step 4 The exact matches are provided as the response for the user
Step 5 The steps are followed until the 100% recall is achieved.

.

Precision

The fraction of the document retrieved should be relevant to the user information needs.

$$\textit{Precision} = \frac{\textit{No of related terms}}{\textit{Total number of retrieved terms}} \qquad (3)$$

Recall

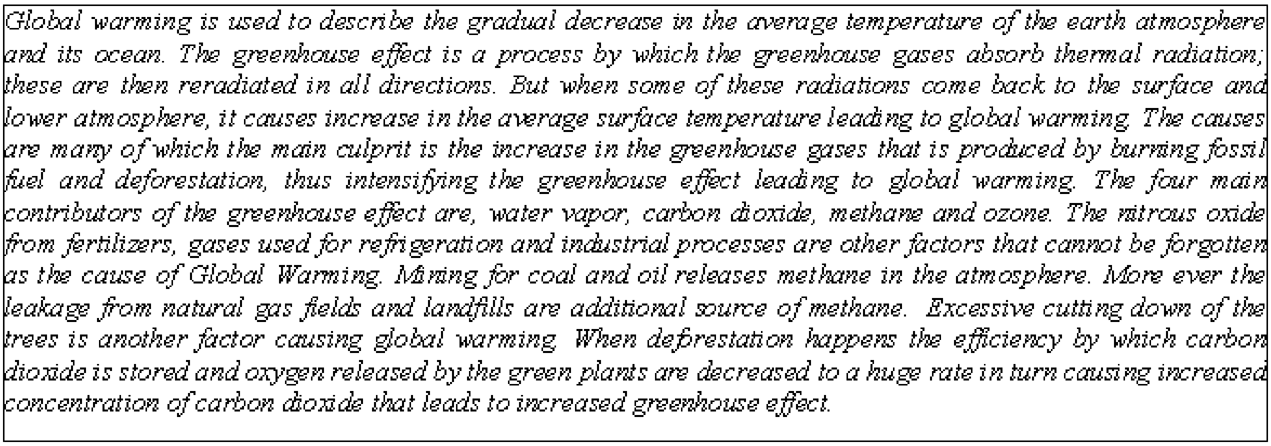
The fraction of the document should be relevant to the query and that are retrieved successfully.

$$\textit{Recall} = \frac{\textit{No of related retrieved terms}}{\textit{Total number of related terms}} \qquad (4)$$

4. Implementation

In our proposed work, we are considering the original document consists of 50-60 words. The .NET framework is used as the front end. The summarized document consists of only 20-25 words. Then according to the user queries the information is retrieved. The future implementation is to compare the retrieved information with the standard tools such as ROUGE.

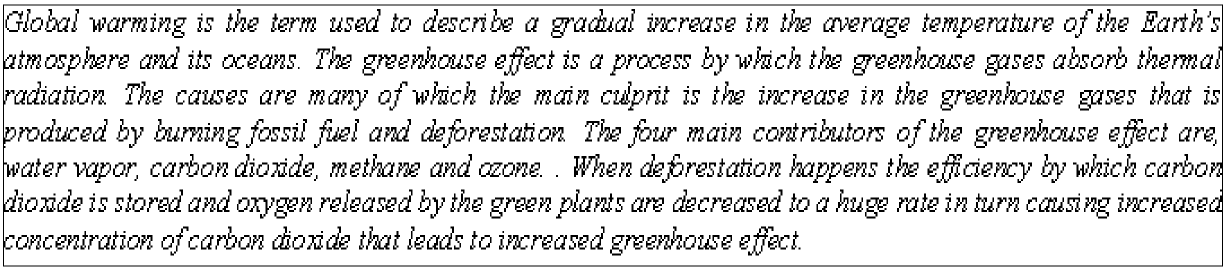
The orginal document gives the information about the global warming



Global warming is used to describe the gradual decrease in the average temperature of the earth atmosphere and its ocean. The greenhouse effect is a process by which the greenhouse gases absorb thermal radiation; these are then reradiated in all directions. But when some of these radiations come back to the surface and lower atmosphere, it causes increase in the average surface temperature leading to global warming. The causes are many of which the main culprit is the increase in the greenhouse gases that is produced by burning fossil fuel and deforestation, thus intensifying the greenhouse effect leading to global warming. The four main contributors of the greenhouse effect are, water vapor, carbon dioxide, methane and ozone. The nitrous oxide from fertilizers, gases used for refrigeration and industrial processes are other factors that cannot be forgotten as the cause of Global Warming. Mining for coal and oil releases methane in the atmosphere. More ever the leakage from natural gas fields and landfills are additional source of methane. Excessive cutting down of the trees is another factor causing global warming. When deforestation happens the efficiency by which carbon dioxide is stored and oxygen released by the green plants are decreased to a huge rate in turn causing increased concentration of carbon dioxide that leads to increased greenhouse effect.

Fig 2 The orginal document

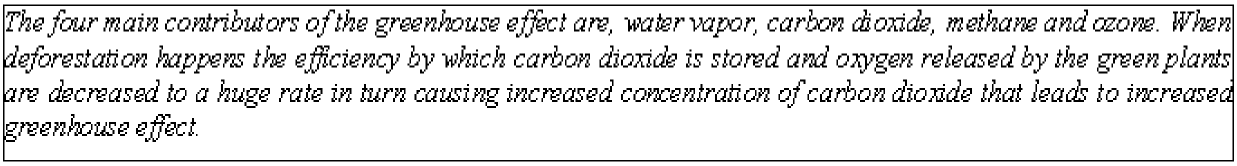
By using Lexical analysis, the indexing weight is calculated for the contemporary terms and finally the document summarization is done.



Global warming is the term used to describe a gradual increase in the average temperature of the Earth's atmosphere and its oceans. The greenhouse effect is a process by which the greenhouse gases absorb thermal radiation. The causes are many of which the main culprit is the increase in the greenhouse gases that is produced by burning fossil fuel and deforestation. The four main contributors of the greenhouse effect are, water vapor, carbon dioxide, methane and ozone. . When deforestation happens the efficiency by which carbon dioxide is stored and oxygen released by the green plants are decreased to a huge rate in turn causing increased concentration of carbon dioxide that leads to increased greenhouse effect.

Fig. 3 The summarized document

By giving the query the user can retrieve the information from the document. If the user wants to know about the green house gases just by giving the query and then by indexing algorithm the information can be retrieved.



The four main contributors of the greenhouse effect are, water vapor, carbon dioxide, methane and ozone. When deforestation happens the efficiency by which carbon dioxide is stored and oxygen released by the green plants are decreased to a huge rate in turn causing increased concentration of carbon dioxide that leads to increased greenhouse effect.

Fig. 4 The information retrieved from the summarized document

5. Conclusion

In this paper, the context based indexing approach is used for retrieving the information from the summarized document. Normally the original document consists of a large amount of information. The user finds more difficult to gather the idea about the content in the original paper. In order to avoid this problem, document summarization is used. The document summarization is the process of collecting the most salient sentences for the original document by dividing the document into contemporary terms and the non-contemporary terms. The content contemporary terms are collected and then the indexing weight is applied for each term. The terms containing the highest indexing weight are extracted and finally the document summarization is done effectively. Then according to user queries the information is retrieved by matching the queries with the summarized document by using indexing algorithm. The sentences which are all matched with the query are given as the responses for the user. The future enhancement is to implement the context-based model for event detection.

REFERENCES

[1] C. Shen and T. Li, “Multi-Document Summarization *via the Minimum Dominating Set,” Proc.23rd Int’l Conf. Computational Linguistics, pp. 984-992,*

[2] X. Wan, “Towards a Unified Approach to Simultaneous Single- Document and Multi-Document Summarizations,” Proc. 23rd Int’l Conf. Computational Linguistics,pp.1137-1145

[3] C.C. Chen and M.C. Chen, “TSCAN: A Content Anatomy Approach to Temporal Topic Summarization,” IEEE Trans. Knowledge and Data Eng., vol. 24, no. 1, pp. 170-183, Jan. 2012.

[4] C.-Y. Lin, G. Cao, J. Gao, and J.-Y. Nie, “An Information-Theoretic Approach to Automatic Evaluation of Summaries,” Proc. Main Conf. Human Language Technology Conf. North Am. Chapter of the Assoc. of Computational Linguistics, pp. 463-470, http://dx.doi.org/ 10.3115/1220835.1220894, 2006.

[5] K. Morita, E.-S. Atlam, M. Fuketra, K. Tsuda, M. Oono, and J.-i. Aoe, “Word Classification and Hierarchy using Co-Occurrence Word Information,” Information Processing and Management, vol. 40, pp. 957-972.

[6] H. Li, “Word Clustering and Disambiguation Based on Co-Occurrence Data,” Nat’l Language Eng., vol. 8, pp. 25-42, Mar. 2002.

[7] Xiaojun Wan Jianwu Yang Jianguo Xia ”Towards an iterative reinforcement approach for simultaneous document summarization and keyword extraction”Nov 2005.

[9] E. Hovy and C.-Y. Lin, “Automated Text Summarization and the Summarist System,” Proc. Workshop Held at Baltimore, Maryland pp. 197-214.

[10] I. Mani, G. Klein, D. House, L. Hirschman, T. Firmin, and B. Sundheim, “Summac: A Text Summarization Evaluation,” Nat’l Language Eng., vol. 8, pp. 43-68,

[11] Pawan Goyal, Laxmidhar Behera “A Context-Based Word Indexing Model for Document Summarization “2013.