Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 5, May 2015, pg.521 – 526

RESEARCH ARTICLE

Optimizing Sentence Scoring Method for Query Based Text Summarization

Twinkle A. Rathod[1], Prof. Nikita D. Patel[2]
Kalol Institute of Technology and Research Center, India[1]
Asst. Prof, computer Department, KITRC, India[2]
twinkle7787@gmail.com [1]; emailtoniki@gmail.com [2]

Abstract— Text summarization is the part of Information Retrieval system which comes under the area of Text Mining. A general format for storing data is text which is easy but unstructured. Text mining deals with the unstructured data and finds the interesting data. Text summary is important now a days for online library system that stores newspapers, books or/and magazine. The user can easily find out their interested data from above mentioned data source. Query based text summarization is process of generation of summary where each sentence in the summary is chosen as per the user given query. To generate a query Based text summary, sentence scoring is most important process at a whole. Statistical and linguistic approaches are followed for sentence scoring. Here to combine both and applying weighted average on each sentence scoring method will improve the results in comparison with simple average of those sentence scoring method.

Key words: text mining, information retrieval, sentence scoring

I. Introduction

Every human stores their data in mostly text format. At every place like government offices, financial company's data are being stored in text format. Infect survey also says that the most data (about 80%) are stored in text format by human beings. So, text mining has large scope to get better and find better solutions. It is quite complex and fuzzy task as it needs to be dealt with unstructured data.

The process of text mining is the extraction of non-trivial and interesting data from the unstructured text. Text mining makes the use of different search techniques, but the difference between searching and text mining is that search method needs a user to know what he or she is looking for, whereas text mining attempts to find information in a pattern which is not known before [1].

Text Summarization comes under the area of information retrieval. It condenses the source text into a shorter version preserving its information content and overall meaning. It is very difficult for human beings to manually summarize large documents of text [2].

Text Summarization is having two main approaches:

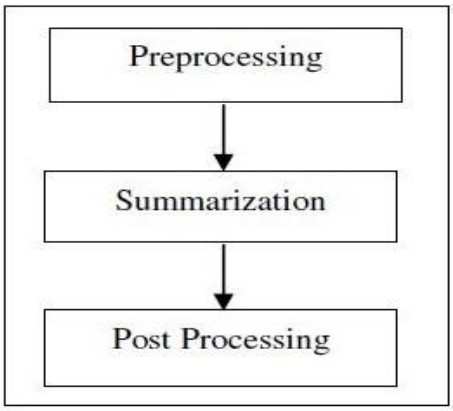
1) Extractive approach
2) Abstractive approach.

Query based text summarization is content based text summarization. It is formed based on user defined keywords or sentences or any numerical data. Query based text summarization can be done using both approaches. 1) statistic 2) linguistics. Statistic techniques are based on structure of the sentences of particular language. Linguistic techniques are related to particular language's grammar.

The main steps of query based text summarization are as follows:
-keyword extraction
- sentence scoring
- sentence ordering
-generating summary
Sentence scoring can be done using statistical techniques and/or linguistic technique. Sentence scoring is very important part as it decides the summary content .The result shows the summary based on used given data which is important for user. The problem is to how to score each sentence such that it will generate meaningful and precise summary in terms of the user given query. And the result should be as per user's concerned topic which is given in terms of query. Query can be word, sentence or any phrase.

II. Background Theory

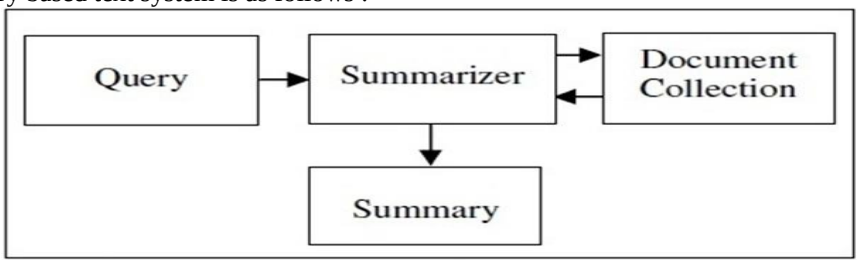
Text Summarization:
Text summarization is the process that takes important data from a source text. Text summarization is very challenging task in information retrieval. . The use of text summarization allows a user to get a sense of the content of full-text, or to know its Information content without reading all sentences within the full-text[6].



The text summarization process works in above figure shown manner. The first stage is preprocessing the given text. Pre Processing is structured representation of the original text. It usually includes: Sentences boundary identification, Stop-Word Elimination, Stemming.
The second step is approach based.
-Extractive summary
-Abstractive summary

Problems with Abstractive Summary[7]:
-The biggest challenge is representation of problem. System's capabilities are constrained by richness of its representations and ability to generate structures.
- Abstractive summarization works on mainly semantic features of a given text. The system can only generate worth summary if having the capability to 'understand' natural language.
-To work with semantics of text is more complex than structural information.

Query Based Text Summarization Method
The generic query based text system is as follows :

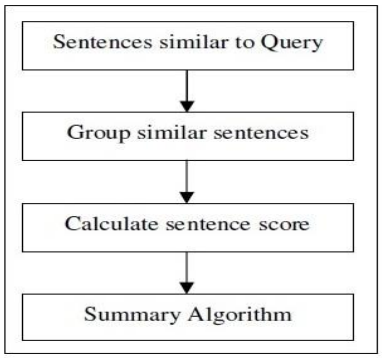


Query Based Text Summarization System

In query based text summarization[10] system, the sentences in a given document are scored based on the frequency counts of terms (words or phrases). The sentences containing the query phrases are given higher scores than the ones containing single query words. Then, the sentences with highest scores are incorporated into the output summary together with their structural context. Portions of text may be extracted from different sections or subsections. The resulting summary is the union of such extracts.

III. Literature Review

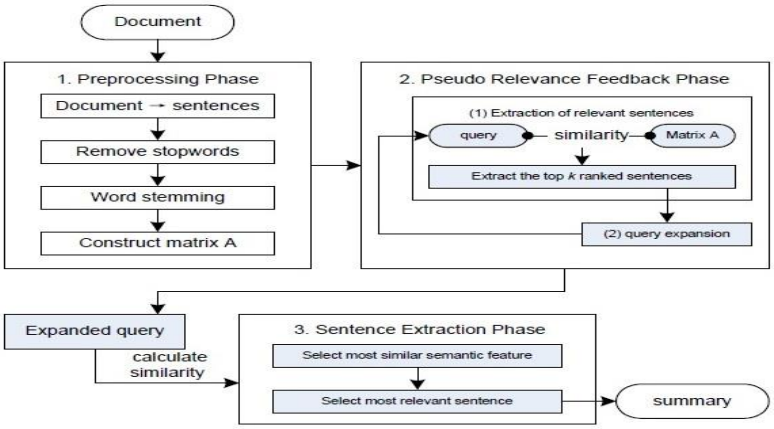
1) Query-Based Summarizer Based on Similarity of Sentences and Word Frequency[6]
The summary is generated based on calculating sentence similarity to the query. The similarity is calculated using cosine similarity score. The steps followed are as follows:



Stages of summarizer[6]

- Sentence similar to query: using cosine similarity sentence similar to query is counted.
- Group similar sentences :sentences are arranged in ascending order based on similarity value then group is formed according to values falling into particular group.
- Calculate sentence score : sentence score is counted based on word weights.
- Summary algorithm :
  1. Compute Word Weight Score
  2. Compute Sentence Score and Sentence Location Score
  3. Calculate Group Score
  4. Arrange groups in ascending order as per group score
  5. From best group pick sentences having maximum sentence score
  6. Delete group and repeat step 5 until each group is processed.

2) User-focused Automatic Document Summarization using Nonnegative Matrix Factorization and Pseudo Relevance Feedback[22]
This paper proposes an automatic document summarization method using the pseudo relevance Feedback (PRF) and the non-negative matrix factorization (NMF) to extract sentences relevant to a user are interesting for user-focused summary.



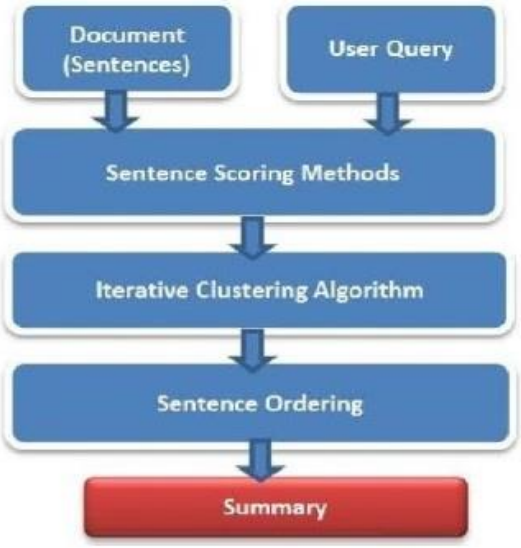
Use Focused Document summarization system[8]

Step:1 Pre-processing Step:
In the preprocessing phase, documents are decomposed into individual sentences, stop-words are removed, and word stemming is performed. Then the term-frequency vectors for all sentences in documents are constructed.

Step:2 Pseudo Relevance Feedback Phase : calculates the cosine similarity between the initial query and a sentence vector by using equation and then selects the top k ranked sentences having the high similarity values.

Step-3 Sentence extraction phase: The sentence extraction phase uses the NMF to extract the sentences for document summary. NMF is to decompose a given matrix A into a nonnegative semantic feature matrix W and a non-negative semantic variable matrix H.
We calculate the similarity between query and semantic feature vectors
-select the semantic feature vector having the highest similarity value.
- We select the semantic variable vector corresponding to the selected semantic feature vector.
-We extract the sentence corresponding to the largest value of semantic variable.
-We repeat these steps until the predefined number of sentences to be summarized is reached.

3) A Hybrid Method For Query Based Automatic Summarization System[7]
A sentence scoring method is defined based on existing sentence scoring methods. It attempts to combine the individual results of these methods to give a better assessment of the relationship between the sentences.



Summarization System[9]

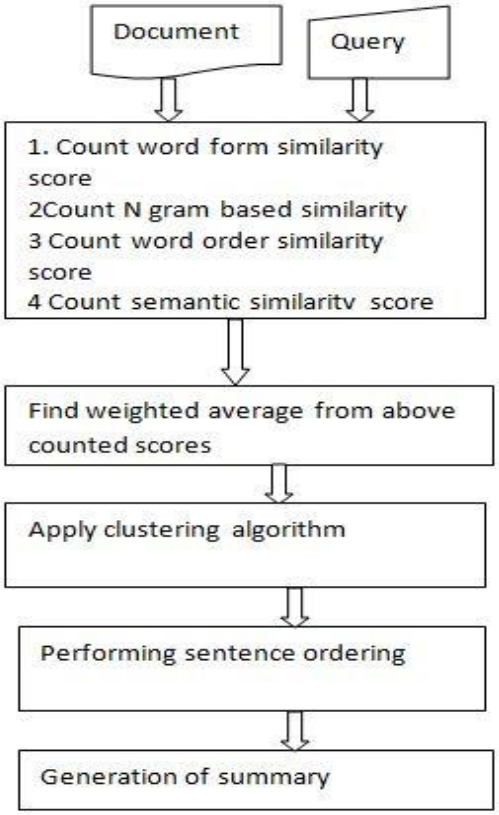
Step 1 : Sentence Scoring Methods
Sentence scoring methods are very important in a document summarization. The efficiency of summarization system mostly depends on sentence scoring method. The main task of the sentence scoring methods is to identify set of sentences which will carry important data in the given document. The scoring methods are also known as "Sentence Similarity Measures".
Statistical and linguistic techniques are used for calculating similarity score.
-Statistical Techniques:
a. Word form similarity
b. N-gram based similarity
c. Word Order Similarity
- Linguistic Techniques:
d. Semantic similarity
Proposed sentence scoring method = ((a+b+c)/3+d)/2
Step-2 Iterative Clustering Algorithm:
After retrieving the important sentences by applying the proposed sentence scoring method, there is a need to check for redundancy among the sentences. One way to remove redundancy is through sentence clustering.
1. The extracted sentences S are arranged in ascending order on the basis of score.
2. First sentence is selected and its similarity is measured with all the other sentence.
3. Sentences having similarity above the threshold are removed from the set S Similarly
4. The procedure is repeated for all the sentences and then the outcome will be a summary without redundancy.

Step 3 Sentence Ordering:
To generate a readable coherent summary it is very important to order the sentences correctly.
For a single document the order of sentence in the original document can be used as order to generate summaries. Alternately the sentences can be presented in descending order of their score.

IV. Methodology

- **Step 1:** Enter the Data
  - Read the document to generate a summary
  - Read the user defined query
- **Step 2:** Sentence Scoring Phase
  - Calculate the word form similarity
  - Calculate the N-gram based similarity
  - Calculate the word order similarity
  - Calculate the semantic similarity
  - Calculate weighted average of above score using below equation
- **Step 3:** Sentence Clustering[24]
  - The extracted sentences *S* are arranged in ascending order on the basis of score
  - First sentence is selected and its similarity is measured with all the other sentences
  - Sentences having similarity above the threshold(50%) are removed from the set S
  - The procedure is repeated for all the sentences
- **Step 4:** Sentence Ordering
  - The sentences are ordered in the order of original document.
- **Step 5:** Generation of Query based text summary.

Flow Of Work:



Conclusion and Future Work
Using hybrid method for sentence scoring and using weighted average of three statistic method and linguistic method the generated query based text summary gives better result. For future work this system can be extended for the multi document query based text summarization.

References

[1]. Miss. Dipti Shyam Charjan;" Review of Text Mining Method: Investigation and Analysis," IRACST – International Journal of Advanced Computing, Engineering and Application (IJACEA), Vol.2, No. 1, 2013

[2]. Vishal Gupta, Gurpreet Singh Lehal" A Survey of Text Summarization Extractive Techniques," JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 2, NO. 3, AUGUST 2010

[3]. Saeedeh Gholamrezazadeh Mohsen Amini Salehi Bahareh Gholamzadeh" A Comprehensive Survey on Text Summarization Systems,"

[4]. Debora Cheney" Text mining newspapers and news content: new trends and research methodologies"ILFA WLIC Singapore, 2013

[5] Rashmi Agrawal, Mridula Batra;" " A Detailed Study on Text Mining Techniques," International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-6, January 2013

[6] A. P. Siva kumar1, Dr. P. Premchand2 and Dr. A. Govardhan3;" Query-Based Summarizer Based on Similarity of Sentences and Word Frequency" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.3, May 2011

[7] Jimmy Lin., "limitation of abstractive text summarization Summarization.", Encyclopedia of Database Systems. Heidelberg, Germany: Springer-Verlag, 2009.

[8] Sun Park; "User-focused Automatic Document Summarization using Nonnegative Matrix Factorization and Pseudo Relevance Feedback", 2009 International Conference on Computer Engineering and Applications IPCSIT vol.2 (2011) © (2011) IACSIT Press, Singapore

[9] RVV Murli Krishna,SY Pavan Kumar, ch. Styandra Reddy;"a hybrid method for query based automatic summarization system", International Journal of Computer Applications (0975 – 8887) Volume 68– No.6, April 2013

[10] Azadeh Zamanifar, Behrouz minaei-Bidgoli and Mohsen Sharifi," A New Hybrid Farsi Text Summarization" Technique Based on Term Co-Occurrence and Conceptual Property of Text ", In Proceedings of Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, IEEE, 635-639, Iran,2008.

