An Effective Approach for News Article Summarization

Shilpi Malhotra
Department of Computer Engineering
YMCA University of Science &Technology
Faridabad, India

Ashutosh Dixit
Department of Computer Engineering
YMCA University of Science &Technology
Faridabad, India

ABSTRACT

Information on the World Wide Web and in other electronic form is increasing tremendously. Therefore there is a need for some form of information compression which can be achieved by various mining tasks like classification, clustering and summarization that help in understanding the information. Large amount of web content is news. News websites are daily overwhelmed with plenty of news articles. This paper presents an effective approach for single document news article summarization to help people obtain the most important information in the shortest time. The proposed approach is query based news article summarization. The results from web based on user query are filtered and refined and then result is directed to user. The technique used for summarization is keyword based extractive summarization. Keywords are the index terms that contain the most important information. The summarization technique identifies different features like thematic terms, named entity, title terms, numbers etc that are relevant to news articles to construct keyword table. This knowledge base is then used to score sentences and then top ranked sentences are presented as summary to the user. For evaluation of summary generated, extrinsic technique by question answering system is used. The purpose of using this evaluation technique is to test if the summary can be used instead of original document while preserving the overall importance of the document i.e. can summary covers all the important information of the document.

General Terms

Extractive Summarization; Extrinsic Evaluation; Sentence Extraction; Sentence Filter;

Keywords

Corpus Builder; Headline Similarity; Keyword Table; Named Entities; Thematic terms

1. INTRODUCTION

With the tremendous increase of digitized information, the mining task has become a crucial tool for aiding and understanding the information. This includes clustering, classification, categorization and summarization. The major challenge is to find relevant information from large amount of data. Summaries are often necessary to enable timely relevancy assessments, information extraction, or information analysis from source material. Text summarization is an effective technique that is used in combination with Information Retrieval and Information filtering systems to save the user time. [1]

Today the size of the repository of information is much larger than one can manage, easily and efficiently. This includes business transactions, news reports, satellite data, digital media, text reports and memos and biological information**.** [2] Moreover in today's life everyone wants to gain more and more in less time. Thus reading long documents and then gaining the insight of the document is not a good idea. It will be more beneficial if one go through the summary of the long document and still gaining the theme or core information present in the document. In this way more and more information can be gathered in less time. Thus the demand for efficient data mining techniques is increasing day by day.

Now-a-days there are plenty of online news websites overwhelmed with news articles. The most important tasks of news engines are Collecting News, News Retrieval, Categorizing Search Result, Summarization and Automatic Event Detection. The quality of each of the tasks depends on the quality of several other tasks. [3] This paper focuses on simple technique to take query from user and receive the ranked news related to the user's query from web. The irrelevant news articles are discarded and user gets refined data. The technique used to produce extractive summary for single news article that carries most important information is keyword extraction technique.

2. RELATED WORK

Sparck Jones [4] defines a summary as, "A reductive transformation of source text to summary text through content reduction by selection and/or generalization on what is important in the source".

Automatic text summarization has been studied for over 50 years now. Luhn [5] in 1958, suggested to weight the sentences of a document based on term filtering and word frequency is carried out (low-frequency terms are removed), sentences are weighted by the significant terms they contained. Automatic text summarization system [6] in 1969, which, in addition to the standard keyword method (i.e., frequency depending weights), also used the pragmatic term method (cue words like hence, finally etc), title term method and location method to determine the sentence weights.

Text Summarization methods can be classified into extractive and abstractive summarization based on the origin of text in the summary. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. An abstractive summarization attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language. It uses linguistic methods to

examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. The input to summarization process can be single or multiple documents. The summary can be indicative that provide an idea of the text, generic that gives overall sense of the text or focused that contain information based on user query. [7, 8, 9]

In addition to the research challenges in developing automatic summarization systems, how good to evaluate their result has emerged as a research issue in itself. With the literature available the evaluation techniques can be divided into intrinsic and extrinsic techniques. Intrinsic evaluation techniques focus on content of the summary. Some of the intrinsic evaluation techniques are recall and precision, DUC (Document Understanding Conference), relative utility and ROUGE (Recall Oriented Understudy for Gisting Evaluation) automatic evaluation. Extrinsic evaluation measures how well do the summary help a user with a task. Extrinsic evaluation tries to find out whether summary can be used instead of document; can the document be used to classify document; can one answer questions by reading the summary. [10, 11]

3. PROPOSED ARCHITECTURE

This paper proposes a scheme for query based single document news article summarization. The proposed architecture (shown in Fig.1) works for news articles retrieved from Web as a result of query terms entered by user.

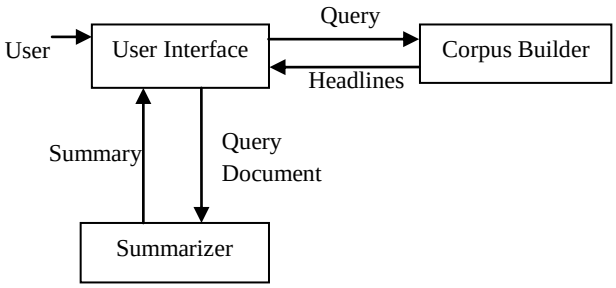

Fig. 1 Architecture of Proposed System

Initially user enters the query for which he/she wants the information. This query is passed to corpus builder that retrieves the resultant news pages from the web, discard the irrelevant news pages and the headlines of relevant news pages are returned to user. User then selects the headline and passes it to summarizer which then applies keyword extraction method to generate summary and return it to user.

The proposed architecture consists of following components:

1) Corpus Builder

2) Summarizer

The working of these component modules is explained below.

3.1 Corpus Builder

This module collects news pages retrieved from the web corresponding to the query terms. The user enters query terms through user interface. This query is passed to news collector that retrieves the resultant news pages from the web. The headlines of the retrieved news pages are tokenized so as to find the headline similarity. All the news pages that are having similarity value above some set threshold say Θ_h are added to the corpus. The similarity between headline h and query q is calculated using (1):

$$Sim(q, h) = \frac{KW(q) \wedge KW(h)}{KW(q) \vee KW(h)} \quad (1)$$

The numerator gives the common terms in headline h and query q whereas denominator gives the total number of terms in headline h and query q. Let there are M news pages retrieved as a result from Web. Out of M news pages N news pages have headline similarity above Θ_h. These N news pages are directly added to corpus. For all other M- N retrieved documents, first four paragraphs are extracted and then the score for each news page is calculated. The score for each document is calculated using (2):

$$Score(NW_i) = \sum_{j=1}^{N} tf_{j,i} \quad (2)$$

Where NW_i is the ith news page, N is the total number of query terms and $tf_{j,i}$ is the term frequency of jth term in ith document. This calculates the frequency of query terms in first four paragraphs. Those with score above predefined threshold say Θ_p are then added to corpus.

The headlines of the news pages added to the corpus is then directed to the user so as to get the feedback. User then selects any headline for which he/she wants summarized information and is passed to module summarizer. The detailed architecture of this module is given in fig 2.

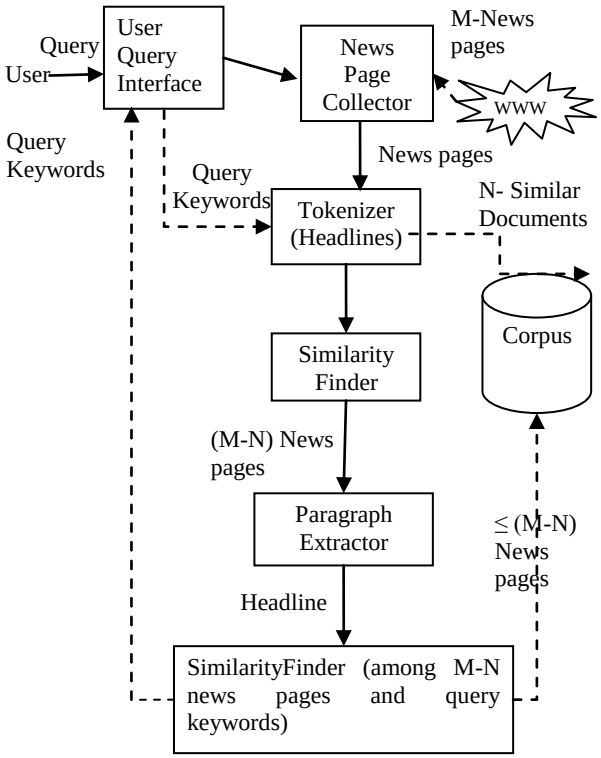

Fig. 2 Architecture of Corpus Builder

Algorithm: Corpus building

Input: Query terms

Step 1: Send query to search engine

Step 2: Perform step 3 to 8 for each retrieved headline

Step 3: Compute Similarity between query and the headline using Sim(q,h)

$$Sim(q, h) = \frac{KW(q) \wedge KW(h)}{KW(q) \vee KW(h)}$$

Step 4: If $Sim(q,h) > \Theta_h$

Step 5: Add headline and its corresponding document to corpus.

Step 6: Else

Step 7: Extract article and score for the article using $Score(NW_i)$

$$Score(NW_i) = \sum_{j=1}^{N} tf_{j,i}$$

Step 8: If $Score(NW_i) > \Theta_p$

Step 9: Add headline and corresponding document to corpus.

Output: List of headlines

Fig. 3 Corpus Builder Algorithm

3.2 Summarizer: News page corresponding to selected headline is pre-processed first. This module takes the document and performs some sort of pre-processing so as to obtain an intermediate representation. Then keywords are extracted and weighted which are then later used by sentence ranker to calculate sentence relevancy. The sentences are then ranked and k top most ranked sentences are presented as summary to user. The summarizer includes: pre-processing, keyword extraction, sentence ranking and sentence filtering. Fig.4 shows the detailed architecture of module summarizer.

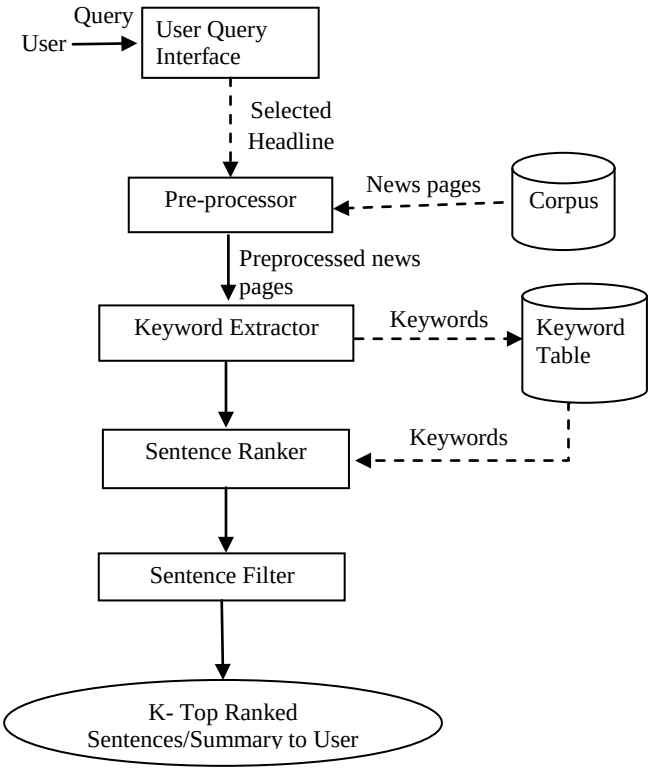

Fig.4 Architecture of Summarizer

The working of different components of summarizer is as follows:

3.1.2.1 Pre-processing: In pre-processing, first main article is extracted from the news page, stopwords are eliminated, and light stemming is performed (i.e. only plural forms are stemmed).

3.1.2.2 Keyword Extraction: Keywords are the index terms that contain most important information. Kaur and Gupta [12] discussed the different approaches to identify keywords. The quality of summary highly depends on keyword extracted which is only possible when several features are combined. The proposed system identifies the keywords using the following approaches:

3.1.2.2.1 Title Word Feature: Title and heading of a news story is often strongly related to its content. Hence words occurring in the title and heading are considered as important indicators for measuring importance of sentences in a document. Title words excluding stopwords form the title glossary for the news story.

3.1.2.2.2 Named Entity: Named entity refers persons, locations and organizations. The named entities are most informative. The occurrence of such entities represents clues of positive relevance of a sentence for the summary, especially in news text. All such entities are identified and added to keyword table with their frequency.

3.1.2.2.3 Nouns: Nouns are considered the conceptual entities in text documents. The noun terms are identified and their frequency is also calculated. Since some of the noun terms especially proper nouns are already added to the keyword table through named entity recognition therefore nouns other than them are identified and added to the keyword table. POS (Parts Of Speech) tagging is performed to identify noun terms.

3.1.2.2.4 Thematic Terms: Statistics provides clues that important sentences are the ones that contain words that occur frequently. Such terms are called thematic words. The term frequency of a term is the number of occurrences of that term in the whole document.10 – 15% most frequent terms out of total terms are considered as keywords and added to keyword table.

3.1.2.2.5 Temporal Expressions: Sentences containing days, weeks, months or time are important in news articles. Therefore add date and time to keyword table.

3.1.2.2.6 Numeral Data: The sentences that contain any sort of numerical data are scored higher than those that do not contain. Add numeral data to keyword table.

3.1.2.2.7 Anchor Text: News articles contain some anchor text very often that are linked to some other page. These terms are those that are relevant not only to query article but also to other news article.

3.1.2.2.8 Location Feature: Certain types of documents have their key meaning in certain parts of it. For example first sentence of a news article is very important as it covers information regarding Who? What? Where? and When? of the story. This feature is not used for keyword construction but the first sentence of the article is included in the summary as mandatory.

The above features are combined to assign weight to each sentence. The steps to construct keyword table is given in fig.5.

Initial Requirement: Preprocessed news article and top thematic terms

Algorithm:

Step 1: Add title terms to Keyword Table and assign weight to them.

Step 2: Perform 3 to 12 for each sentence

Step 3: Perform 4 to 7 for each named entity NE identified

Step 4: if NE already in Keyword table

Step 5: update its weight

Step 6: else

Step 7: make a new entry and assign weight to its term frequency

Step 8: Perform 9 to 12 for each noun phrase NP (other than proper noun) and cardinal Number CD identified

Step 9: if NP or CD already in Keyword table

Step 10: update its weight

Step 11: else

Step 12: make a new entry and assign weight to its term frequency

Step 13: Perform 14 to 17 for each thematic term TT identified

Step 14: if TT already in Keyword table

Step 15: do nothing

Step 16: else

Step 17: make a new entry and assign weight to its term frequency

Step 18: Perform 19 to 22 for each anchor text AT identified

Step 19: if AT already in Keyword table

Step 20: update its weight

Step 21: else

Step 22: make a new entry and assign weight to its term frequency

Output: Keyword Table

Fig. 5 Keyword Table Construction

3.1.3 Sentence Ranker: Sentence ranker computes the score for each sentence based on these features so that the sentence that is most informative is included in the summary. Score for each sentence S_i can be calculated using (3):

$$SScore(S_i) = \sum_{j=1}^{N} W_j * KW_j \quad (3)$$

Where KW_j is the jth keyword in sentence S_i and W_j is the weight of the keyword. N is the number of keywords present in the sentence. Weight for each term is taken as frequency of keyword in news article. The first sentence of the article is given highest rank.

3.1.4 Sentence Filter: After ranking the sentence there is a need to eliminate redundant data to make our summary more concise. Ranjna Gupta et. al. [13] finds the similarity between two documents D_i and D_j using (4) and (5). These formulas are used to find the similarity between two sentences S_i and S_j

$$SenSim(S_i, S_j) = \frac{\sum_{k=1}^{n} KW_{k,ti} * KW_{k,tj}}{\text{Length of } S_i * \text{Length of } S_j} \quad (4)$$

where t_i and t_j represent the terms in sentence S_i and S_j. The numerator gives the summation of products of term frequencies of common keywords (which is n in number) in S_i and S_j. The length of the sentence S_i can be calculated by (5):

$$\text{Length of sentence } S_i = \sqrt{\sum_{k=1}^{n} w_{i,k}^2} \quad (5)$$

where k represents the tokens in sentence S_i. The length is calculated by taking the square root of summation of products of term frequencies of tokens/keywords in sentence S_i. If the value of sentence similarity of sentences S_i and S_j is above threshold then one with lower score value is discarded. After elimination of redundant information, few top ranked sentences are extracted depending on the specified summary limit and present them as summary to the user.

4. PERFORMANCE EVALUATION

The results of experiment are presented in this section. A database was created consisting of 50 news pages. A query was fired which was matched against headlines of news pages. Based on headline similarity value 15 pages were selected having headline similarity value above 90% (Θ_h). The score was calculated for other 35 news pages by extracting first four paragraphs. Out of these, three pages are further added to corpus with score above 75% (Θ_p). Therefore headlines of 18 news pages are then presented to user. User selected a headline whose document was then passed to summarizer which generates summary using different features and presented to user.

For the evaluation of summary generated, one of the techniques used is extrinsic evaluation by question answering system. A set of 5 questions were prepared for a news article that can be answered by reading article. Five team of two people each was taken and given the prepared questionnaire and the summary generated by proposed system and existing system and they were asked to answer the questions which are then evaluated to find out the quality of the summary. The purpose of using this evaluation technique is to test if the summary can be used instead of original document while preserving the overall importance of the document i.e. can summary covers all the important information of the document. For this purpose four teams viz. Team A , Team B, Team C and Team D were taken. For the comparison purpose three online summarizers namely summarizer summry[14], Freesummarizer[15], and tools4noobs[16] were used. Same document was passed to each of the online summarizer and summary from each online summarizer was generated which was given to Team A, Team B and Team C respectively. Team D was given summary generated by proposed system. The length of the summary used was 30% of the length of the original document. Along with the summary each team was given same set of five questions and asked to answer them by reading summary only. The results are then compared based on the number of questions correctly answered by each team.

Table 1. comparison of existing and proposed system

Ques. No.	Team A	Team B	Team C	Team D
1	2	2	2	2
2	1	1	1	0
3	1	0	0	1
4	0	0	0	0
5	0	0	1	3
Total	4	3	4	6
Hit	0.44	0.33	0.44	**0.66**

The results of all the systems are given in Table 1. Hit ratio is equal to obtained points divided by maximum points. The results shows that the proposed technique that are based on keywords which in turn based on named entities and nouns shows better results in comparison to other existing systems.

5. CONCLUSION

In the proposed work, the results of a query are further refined so that only relevant pages are directed to user and all other irrelevant news pages are discarded. This has significantly reduces time as all other documents that are unimportant with respect to query are never presented. The domain specific features are used to extract keywords that comprise the theme and weighted. The features are combined so as to boost the accuracy of automatic generated summary. The results showed that the outcome of proposed technique has comparable results with respect to its quality. In this way the technique significantly reduces time both in searching the information among the results returned by search engine and by avoiding reading through whole article.

6. FUTURE WORK

The above technique is implemented for single document. However the extension of this approach can be the aggregation of news article from multiple sources and then producing a concise summary of multiple documents. Some techniques need to be developed that address the challenges of extractive summary that includes proper decomposition of long sentences. Another problem that needs to be addressed with multi document summarization can be understood using an example. Suppose there are two document d1 and d2. The sentence S_i of d1 ranked higher and is included in the summary. Now some sentence S_j of d2 is next higher ranked needs to be included in the summary but S_j talks about something else say S_i talks about Mr. X and S_j talks about Mr. Y. This is the problem with extractive summary in multidocument summarization. This needs to be resolved.

7. REFERENCES

[1] Maybury, Mark T. Merlino, A.E., "Multimedia summaries of broadcast news", Intelligent Information Systems, 1997. IIS '97. Proceedings, vol., no., pp.442, 449, 8-10 Dec1997

[2] Osmar R. Zaiane, "Principles of Knowledge Discovery in Data Bases". Department of Computer Science, University of Alberta.1999 CMPUT690

[3] H Sayyadi, S Salehi, H AbolHassani, "Survey on News Mining Tasks". Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 2007

[4] Sparck Jones, "Automatic summarizing: factors and directions". In Inderjeet Mani and Mark Marbury, editors, Advances in Automatic Text Summarization.MIT Press, 1999

[5] Luhn, H., P. "The Automatic Creation of Literature Abstracts". In Inderjeet Mani and Mark Marbury, editors, Advances in Automatic Text Summarization. MIT Press, 1999.

[6] Edmundson, H. P. "New Methods in Automatic Extracting". In Inderjeet Mani and Mark Marbury, editors, Advances in Automatic Text Summarization. MIT Press, 1999

[7] Vishal Gupta and Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques." Journal Of Emerging Technologies In Web Intelligence, Vol. 2, No. 3, August 2010

[8] Pranitha Reddy and R C Balabantaray, "Improvisation of the Document Summarization by combining the IR techniques with "Code-Quantity and Attention" Linguistic Principles". International Journal of Engineering and Innovative Technology (IJEIT) Volume 1, Issue 5, May 2012

[9] Y. Surendranadha Reddy and Dr. A.P. Siva Kumar, "An Efficient Approach for Web document summarization by Sentence Ranking". International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 7, July 2012.

[10] Nenkova, A. "Summarization evaluation for text and speech: issues and approaches". In: INTERSPEECH-2006, paper 2079-WedlWeS. 1. (2006)

[11] Josef Steinberger, Karel Je˘zek ,"Evaluation measures for text summarization". Computing and Informatics, Vol. 28, 2009, 1001–1026, V 2009-Mar-2

[12] Vishal Gupta, Jasmeen Kaur , "Effective Approaches For Extraction Of Keywords". IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6, November 2010

[13] Ranjna Gupta, Neelam Duhan , A.K. Sharma and Neha Aggarwal, "Query Based Duplicate Data Detection on WWW" . International Journal on Computer Science and Engineering Vol. 02, No. 04, 2010

[14] http://www.smmry.com -"Online Automatic Text Summarization Tool".

[15] http://www.freesummarizer.com -"Online Automatic Text Summarization Tool".

[16] http://www.tools4noobs. -"Online Automatic Text Summarization Tool".