International Journal of Research in Advent Technology, Vol.7, No.11, November 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

An Effective Approach for Online News Extraction and Summarization for a Single Phase

Senjuthi Bhattacharjee, Asma Joshita Trisha, and Sabrina Akter

Abstract— Online newspaper plays an important role for the development of world. But it consists of several types of labels, titles and links. As online newspapers are collection of variety of newspaper, it is often much more difficult to extract and summarize the news. To improve the accuracy a new algorithm is introduced here based on web extraction and summarization. Firstly, the news from newspapers are extracted which are related to the topic. If different types of news are found about the same topic then it has distinguished. Then a summarization-based algorithm has proposed to summarize the news. Basically, term frequency has used for summarization and evaluate it along with several newspapers' contents. Various forms of words are also compared such as Noun, Adjective, Adverb etc. So that the term frequency can be counted more accurately. It will be very helpful for a user who wants to find out very specific news from the newspapers.

Keywords— Extraction, online news, precision, sentence scoring, summary, term frequency.

I. INTRODUCTION

Information retrieval is the term that specifies extraction of relevant information from various documents. Information retrieval can be done in different ways. Web data extraction is one of them. Data contained in websites (newspaper) is increasing exponentially. But much of this information cannot be used by other applications. As most of the web data will be in XML format, it will solve the problem in future.

But now this is not the case and information in web have to be retrieved efficiently. So, their emerged a new source of information retrieval technique which is extraction of web data. It is a process through which data can be extracted from web without loss of information. Web data is in semi-structured format. To extract data from web, it is necessary to analyze each word and tag found in the particular website.

Present usability of the online news largely depends on news summarization(web). Tailoring of the content of Web documents to match specific displays through web document summarization in an accessibility purpose, mainly range from snippet generation by search engines, (e.g. for blind people). To summarize automatically, plain text document is used.

In an HTML document there are many elements like as pictures, which cannot be summarized and it is difficult to distinguish the relevant information among many news. In recent years, many applications are introduced which particularly works with the content of a HTML document. Here the context of the document has used where information is retrieved from all the documents linking to it.

Online news Summarization is a technique that search newspaper for specific query and returns a compact summary for a given newspaper to representing its main content. Here the main purpose is to generate a compatible summary which are as good as the summaries done by a person.

Textual snippet is the most widespread search-based summarization (Zhanying He et al., 2013). When a user submitted a query, web search engine provides the reference for sequence of top-k documents. Each document contains a title, a snippet, a URL. When there is less time for browsing the site, web summary helps user to get idea about the content of the page. This extracts the sentences which are more significant from a web page and generates a summary to the user. The web includes different kind of information like text, images, video and audio. So, it is necessary to extract relevant result. The good web page summary must be a clear, a simple guide what is on the page.

There are two types of summary such as an abstract and an extract. When the summary consists of remarkable text units selected from the input then it is called extract summary (N.Moratanch and S.Chitrakala, 2017). An abstract is a brief summary of a definite subject, which are generated by computing the noticeable units selected from an input. Text units which are not present in input text can also be included in abstract summary (N. Moratanch and S. Chitrakala, 2016).

II. RELATED WORKS

A well-known method is the centroid-based method (Xindong Wu et al., 2011), in this method, TFID feature is used for calculating the sentence score. For each single feature, the score is calculated and then combine it for the whole sentence. To extract and summarize online newspaper for a single phrase it is required to categorize the news firstly. Then summarization of the particular portion is done. There is a approach named Conditional Random Fields (CRF) based

Manuscript revised November 23, 2019 and published on December 03, 2019
***Senjuthi Bhattacharjee**, Lecturer, Dept. of Computer Science & Engineering, Premier University,* Chattogram , *Bangladesh.*
***Asma Joshita Trisha,** Lecturer, Dept. of Computer Science & Engineering, Premier University,* Chattogram , *Bangladesh.*
***Sabrina Akter**, Student, Dept. of Computer Science & Engineering, Chittagong University of Engineering & Technology, Chattogram, Bangladesh.*

framework to treat the summarization task as a sequence labeling problem (Dou Shen et al., 2007). The sentences which has highest scores are extracted in extraction-based summarization (Xiaojun Wan et al., 2007). There are some approaches which mainly combines several sentence features (Minqing Hu and Bing Liu, 2006). Now-a-days there are various extraction-based approach for web classification/categorization (Ioannis Antonellis et al., 2006) and summarization (Furu Wei et al., 2008). Sentence redundancy is a big obstacle for summary sentences. To remove redundancy between summary sentences, The MMR algorithm (Mohammad Al Hasan, 2009) is also another popular approach. The Frequent Pattern Mining (FPM) algorithms (Mohammad Al Hasan, 2009) is also used to calculate the complex features, such as set, sequence, tree, graph, etc. But large output set size causes lacking of interpretability, and that's why potentiality of this approach is very low comparing to another algorithm.

An online newspaper generally contains a variety of information cantered around a main title. To get the summarized news for a single phrase, section-based categorization (Giuseppe Attardi et al., 1999) is more workable than other ways. For getting the filter news from various news there can be used K-nearest algorithm and for getting the summarized news there can be used pattern mining or used term frequency.

III. PROPOSED METHOD

Online newspaper contains various types of news. They show the details of news. Now day's readers don't have such time to read all the news. They want to save their time. In this project the user only put a keyword for knowing the news which related with the keyword by extracting news. They also can know the compact news which can cover the all newspaper. People can also know the previous news.

A. Architecture of proposed method

The methodology or architecture of the proposed method is discussed below:

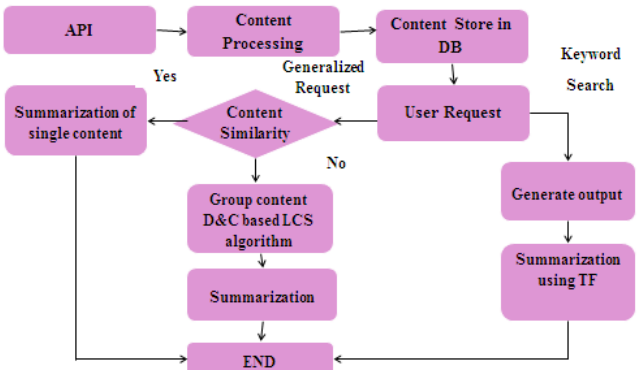

Fig. 1. Architecture of proposed method

B. Step by step description of proposed method

This section gives an analytical description of the system architecture given in previous parts.

B.1. Initialization and Connection

In the initialization and connection module, at first, the web pages of each website are stored in separate files. Then each of these pages will be connected using URL. A table is created in news database for each website having news no, name, date, Headline, description.

B.2. News Extraction

The most important part of this method is news extraction (Y. Sankarasubramaniam et al., 2014). For that at first the input newspapers are taken. Then the keywords will be given as input. Matching the keywords with database contents for extraction. After matching news contents with database contents, the news will be extracted. So, every news is separated in topic wise. For A single domain or phase, different news can be gathered. First, the news of same domain is collected. Then the news in different parts are divided. For cricket much news are found. Such as T-20, One day, Test match etc. Here, the desired news for a particular phrase can be also found. The divide and conquer approach are followed for similar text matching for extraction of news.

- *Divide and Conquer*

In computer science, divide and conquer (D&C) is a method, in which the whole problem is divided into several sub segments and then the whole system is combined to get the solution of the original problem.

- *Similar text matching*

In this method, the query string uses a parameter, which divides the string into low frequency and high frequency group. The low frequency of a group is mainly the more important terms of the bulk of the query, while the high frequency group is the not much important terms is used only for scoring, not for matching.

B.3. News Summarization

The most important part of this method is summarization (J. Goldstein et al., 1999). Here, the extracted news has summarized about the input phrase. In this part, first of all at least two extract news of related phrase has taken from several newspaper. Then every sentence will be checked or compared of this news. In the case of similar sentence, it will take similar sentence at once from both news. The sentences don't be repeated. Then it will summarize the news. Then the process will check, whether there any extract news for summarized. If it is "Yes" then the new news and summary of previous news are summarized by continuing this process. If it is "No", then it will succeed to get the desired output summary. Summarization will be done in using term

frequency. For that some conditions will be applied on the method.

- *Term Frequency*

The importance of a word to a document in a collection or corpus (Xindong Wu et al., 2011) is calculated by term frequency which is a numerical term. It is mainly used to retrieve information retrieve and for text mining. The number of times a word appears in the document, the value of term frequency increases proportionally. It mainly helps to control the common words.

- *Process of summarization using Term Frequency*

Steps of Summarization:
Step 1: First take input Bangla documents as text file.
Step2: In this step tokenized the sentences of input documents and punctuation character, single word, digits are removed from the original Bangla text.
Step 3: Replace each word with common synonym for counting keyword frequency.
Step 4: In this step sort the total term frequency (TTF) in descending order.

Step 5: Compute the score SC_{kj} the kth sentence of the jth document by summing up TTF_i of m number of words in that sentence.

$$SC_{kj} = \sum_{i=1}^{m} (T - n + 1) * TTF_i$$

Step 6: Here all sentence is scored as decreasing order and take only high score sentences that represent the most important sentences in the given documents.
Step 7: Here all sentence is scored as decreasing order and take only high score sentences that represent the most important sentences in the given documents.

IV. RESULT

The main goal of this system is to develop an automatic news extractor and summarizer (Vishal Gupta and Gurpreet Singh, 2013). In this chapter the total implementation process has explained. This chapter also contain a brief description of experimental tools.

A. Tool used for Development

The Tools that are used to develop this method —

- ✓ **Windows 7 Operating System**
- ✓ **Xampp**

B. User Interface

The Interface enables the user to enter the Home Page. There are three sections in home page. 1st section shows all news. It contains all the news in database. Another section is search and the last section is summary.

Fig. 2. Home page

Here, Fig. 3 shows that if user click all news, they get to know the know the all the news which are stored in database for a particular date.

Fig. 3. Output of all news

Fig. 4 describe that of user want to search any keyword for particular news, they get that news if the news available in database, else it shows “no found”.

Fig 4. Output of search news

Fig. 5 shows that if user want to summarize news for a particular topic or every topic, they can get that by using that option.

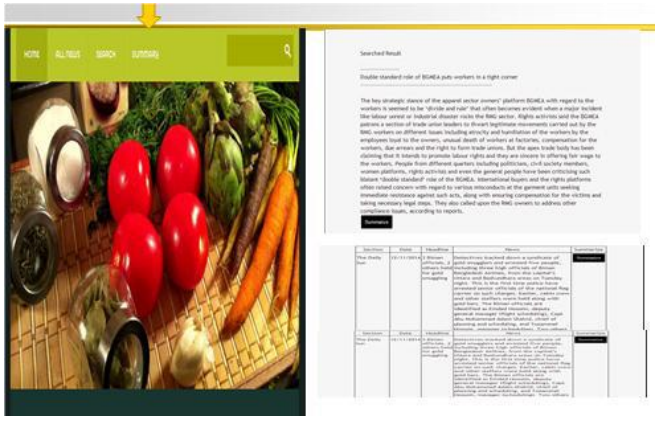

Fig. 5. Options of summary

Here, Fig. 6 shows the desire summary of user which gives the brief news of related news.

Fig 6. Output of summary

C. Experiment Setup

The system retrieves many news from "The Daily Sun", "Bangladesh Independence", "Prothom Alo" (English version) in November. This section contains some experimental results that have been done during experiment. In the following example. A user wants to know about BGMEA. So, System extract the news which is related to BGMEA.

Fig. 7. How to search a keyword

This is the extracting part of this experiment. If user want the summary, he can get that.

News:

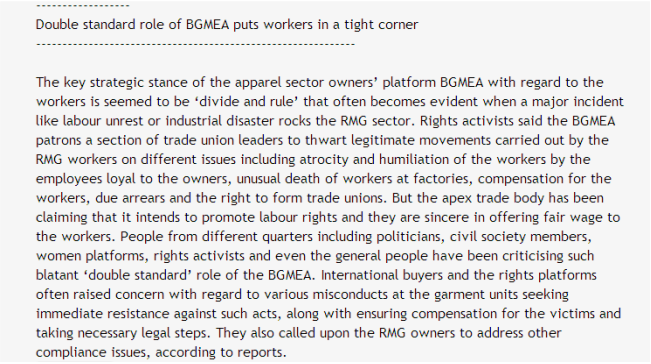

Fig. 8. Input news

D. Term Frequency & Total Term Frequency Count

Most frequent words in the text are the keywords. How many tomes a word appears in the text is counted by the term frequency. Now concatenate each document as a cluster to get total term frequency. Total term frequency is calculated by summing up the term frequency from every document. Sentences with the keywords score higher than those of with fewer keywords. For distinguishing the importance of keyword, the keywords are multiplied which are positioned in higher of the sorted total term frequency value. Table 1 shows the calculation of the occurrence of the keywords.

TABLE I. TERM FREQUENCY OF WORDS

Words	TF	Constant
Worker	8	10
Sector	5	9
BGMEA	4	8
Labour	4	7
RMG	3	6
Owner	3	5
Rule	2	4
Platform	2	3
Industry	2	2
Divide	1	1

E. Sentence Score Generation

Scoring is used to decide on the significance of each line in the documents. Here at most ten sentences are collected for the initial summarized content. The sentence score relies on the word score, which is Total Term Frequency. Final sentence score is the summation of Total Term Frequency.

- **Score of Sentence 1:**

32+15+45+80+1+8+28+4+18+45= 276

- **Score of Sentence 2:**

32+45+36+80+80= 273

- **Score of Sentence 3:**

8+ 28+ +15 = 51

- **Score of Sentence 4:**

28+8=108

Summary:

Fig. 9. Obtained summary

In this summary, it can be observed that most important sentence is obtained high score. The table is given below**,**

TABLE II. SCORE OF SENTENCES IN SUMMARY

No. Sentence in Summary	Score
1st Sentence	32+15+45+80+1+8+28+4+18+45= 276
2nd Sentence	32+45+36+80+80= 273
3rd Sentence	28+80= 108

F. Performance Comparison of the System

To evaluate the system, 7 news sets from different newspapers have gathered. Summarization evaluation methods can divide into two categories: intrinsic and extrinsic (Inderjeet Mani and Mark T. Maybury, 1999).

- ✓ the quality of summaries directly (e.g., by com-paring them to ideal summaries) is measured by the Intrinsic evaluation.
- ✓ how good the summaries help in performing a particular task is measured by extrinsic method.

The system has evaluated in this way-

Compute Intrinsic Measures: Precision, Recall, F-Score and Document Similarity.

TABLE III. INTRINSIC PERFORMANCE ANALYSIS

News no.	Precision	Recall	Fscore
1	0.75	0.6	0.67
2	0.82	0.625	0.71
3	0.6	0.75	0.67
4	0.67	0.5	0.56
5	1	0.45	0.62
6	0.8	0.5	0.61
7	0.7	0.8	0.69

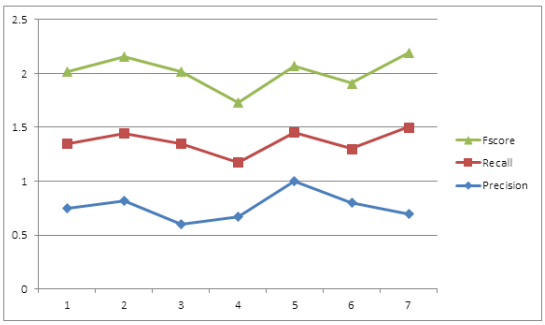

Fig. 10. Intrinsic Performance Analysis Graph

V. CONCLUSION

In this paper, a method has proposed to extract and summarize online newspapers (English) using basic statistical and data mining approaches. Here, challenges have taken for saving times and solving relevancy. Also, the extractive summarization has done more easily and concisely. This work will narrow down the search space for the researchers and thereby save time providing the summary of various news. Moreover, as the methodology followed in this approach is a generic. In future, it can be extended for other newspapers of another languages. In this report, only online newspaper has considered as an isolated document.

VI. REFERENCES

[1] Zhanying He,Chun Chen, Jiajun Bu , Can Wang and Lijun Zhang., " *Document summarization based on data reconstruction."*.Zhejiang Provincial Key Laboratory of Service Robot, College of Computer Science, 2013.

[2] N.Moratanch and S.Chitrakala, "*A Survey on Extractive Text Summarization*", IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP-2017).

[3] N. Moratanch and S. Chitrakala, "A survey on abstractive text summarization," International Conference on Circuit, Power and Computing Technologies (ICCPCT) 2016, International Conference on. IEEE, 2016, pp. 1-7.

[4] Xindong Wu, Fei Xie, Gongqing Wu, Wei Ding. *"Personalized News Filtering and Summarization on the Web", IEEE 23rd International Conference on Tools with Artificial Intelligence, 2011.*

[5] Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. *"Document summarization using conditional random fields", In Proceedings of IJCAI-07.*
[6] Xiaojun Wan, Jianwu Yang, and Jianguo Xiao, *"Manifold-Ranking Based Topic Focused Multi-Document Summarization"*, IJCAI 7 (2007), 2903–2908, 2007.
[7] Minqing Hu and Bing Liu, "*Opinion Extraction and Summarization on the Web",* Department of Computer Science,University of Illinois at Chicago,851 South Morgan Street, Chicago, IL 60607-7053, 2006.
[8] Ioannis Antonellis, Christos Bouras, and Vassilis Poulopoulos "*Personalized News Categorization Through Scalable Text Classification*" Research Academic Computer Technology 36 Institute N. Kazantzaki, University Campus, bGR-26500 Patras, Greece, Computer Engineering and Informatics Department, University of Patras, GR-26500 Patras, Greece, 2006.
[9] Furu Wei, Wenjie Li, Qin Lu and Yanxiang He. "*Query-sensitive mutual reinforcement chain and its application in query-oriented multi-document summarization.* "In Proceedings of SIGIR-08.
[10] Mohammad Al Hasan, "*Summarization in Pattern Mining"*, Encyclopedia of Data Warehousing and Mining, Second Edition, pp.1877-1883, 2009.
[11] Giuseppe Attardi, Antonio Gullì, Fabrizio Sebastiani *"Automatic Web Page Categorization by Link and Context Analysis".* Dipartimento di Informatica, Università di Pisa, Pisa, Italy, 1999.
[12] Y. Sankarasubramaniam, K. Ramanathan, and S. Ghosh, "Text summarization using wikipedia," Information Processing & Management, vol. 50, no. 3, pp. 443-461, 2014.
[13] J. Goldstein, M. Kantrowitz, V. Mittal and J. Carbonell. *"Summarizing Text Documents: Sentence Selection and Evaluation" Metrics.Proceedings of ACM SIGIR-99.*
[14] Vishal Gupta and Gurpreet Singh Lehal, *"Automatic Text Summarization System for Punjabi Language"*, Journal of Emerging Technologies in Web Intelligence 5, 3(2013), 257–271, 2013.
[15] Inderjeet Mani and Mark T. Maybury, "Advances in Automatic Text Summarization", 1999.

AUTHORS PROFILE

Senjuthi Bhattacharjee, B.Sc. in Computer Science & Engineering, Chittagong University of Engineering & Technology, Chattogram, Bangladesh. *Lecturer, Dept. of Computer Science & Engineering, Premier University,* Chattogram , *Bangladesh. (from: January 2016 to Present)*

Asma Joshita Trisha, B.Sc. in Computer Science & Engineering, University of Chittagong, Chattogram, Bangladesh. M.Sc. in Computer Science & Engineering, University of Chittagong, Chattogram, Bangladesh. *Lecturer, Dept. of Computer Science & Engineering, Premier University,* Chattogram, *Bangladesh.* (from: January 2016 to Present)

Sabrina Akter, B.Sc. in Computer Science & Engineering, Chittagong University of Engineering & Technology, Chattogram, Bangladesh.