International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

eStep: A Novel Method for Semantic Text Summarization with Web-based Big Data

Sufal Das

Abstract*: *Text summarization plays an important role in analysis of large set of data. It can be use in online text analysis and knowledge representation. Semantic text summarization plays a vital role to handle big data as data is in very large size, dynamic in nature and heterogeneity. In this paper I have proposed a novel model of knowledge-based semantic analysis for text summarization of web-based dynamic text data with help of FP-tree (Frequent Pattern tree). This model is free from ontology to find out semantic representation. The model consists of two phases. In the first phase benchmark web text data in terrorism domain is collected for construction of domain knowledge representation using FP-tree. Preprocessing is performed to reduce size and handle synonyms. In the second phase, Online articles/news are collected from different sources. Then using the domain knowledge representation, the summary of the web based large text data is extracted.

Keywords* : *Automatic Extract, Semantic Analysis, Summery Evaluation, Text Mining, Text Summarization.

I. INTRODUCTION

Due to rapid growth in communication technology, people can access internet easily. The Internet has facilitated different types of information sharing systems, including the Web. Huge amount of communication is through this media with different corners of the world. Nowadays all sort of information are available in e-world. But main problem is that this information is hidden with large junk as well as distributed environment.

Text summarization is a method to find out useful information from huge collection of text data. Text summarization plays an important role in analysis of large set of data. It can be use in online text analysis and knowledge representation. Semantic text summarization plays a vital role to handle big data as data is in very large size, dynamic in nature and heterogeneity.

Automatic text summarization is based on statistical, linguistically and heuristic methods where the summarization system calculates how often certain key words . The key words belong to the so called open class words. The extraction summarization system calculates the frequency of the key words in the text, which sentences they are present in, and where these sentences are in the text. It considers if the text is tagged with bold text tag, first paragraph tag or numerical values. All this information is compiled and used to summarize the original text.

The main aim here is to design a abstract text summarization for web based text data available on the web without generating ontology from domain. Here we have collected data from the website Global Terrorism Database which is huge in size. Basically the database is dynamic as new incidents in terrorist activities are occurred time by time. Those datasets are preprocessed to remove stop words and synonyms and to get stem words. To handle dynamic data, FP-tree is considered which is generated from the preprocessed text data. For new entry in the dataset, existing FP tree is updated only. Here I have considered that keywords of text appear mostly. Then I have collected many articles which are available in different sources. These collected articles are also preprocessed like the previous data set. This preprocessing removes the stop words and synonyms and produce stem words. Then I have measured similarity index with respect to already generated FP-tree to create a summary of the articles using concept matching. If one paragraph of an article has more than threshold value for similarity then the paragraph is added in summery otherwise rejected.

II. BACKGROUND STUDY

A. Big Data Concept

Author Data is being collected every second in our day to day life. Mainly different sectors like communication, corporate, medical, social network etc. are generation tremendous amount of data regularly. Due to that, all types of researcher have to consider this large volume datasets. Since, data is being generated continuously from different sources; researchers have to also handle dynamic and heterogeneous characteristics of data. Thus data sets become so large and complex that traditional database system and its applications are inadequate.

The Big data [3, 5] can be described in 4 V's: Volume, Velocity, Veracity and Variety.

Volume: It relates to the quantity of data that is generated everyday in large scale and its size is increasing continuously. It refers the large size of input data that can't be handled by traditional database system.

Velocity: Since data is being generated continuously, end users have to consider for online data sets. Velocity refers the characteristic of data with speed of generation of data as well as processing of that data to meet the demands and challenges which is related to the path of growth and development.

Manuscript published on 30 September 2019
* Correspondence Author
Sufal Das*, Information Technology, North-Eastern Hill University, Shillong, India. Email: sufal.das@gmail.com

Veracity: It is very important to consider relevant information from very large datasets. Veracity refers to the quality of the data being captured can vary significantly. Accuracy of analysis depends on the veracity of the input data.

Variety: The term 'variety' includes structured data like tabular data (databases), transactions etc. and unstructured and semi-structured data like hierarchical data, documents, e-mail, video, images, audio etc..

B. Semantic Analysis

Semantic usually means the meaning expressed by a word in a language. Like syntactic there are many approaches are there for semantic exploration of a language. So we can say semantic analysis as an analysis of meaning of elements present in a language. Semantics and its representation include many fields. It cover many tasks like finding synonyms, word sense disambiguation, constructing question-answering systems, translating from one natural language to another, populating base of knowledge [24]. Before doing semantic analysis one should do syntactical analysis. In natural language processing, semantic analysis is in many ways such semantic vector extraction for text summarization, finding semantic similarity between concepts and words.[5].

C. FP-tree

Frequent Item Set (FIS) mining is an essential part of many Machine Learning algorithms.

Let T be a list of n transactions $[t_1, t_2, ..., t_n]$. Each transaction ti contains a list of k_t items $[a_{i1}, a_{2i}, ..., a_{ik}]$. So we have T = [$t_1 = [a_{11}, a_{12}, ... a_{1k}]$, $t_2 = [a_{21}, a_{22}, ... a_{2k}]$, ..., $t_k = [a_{t1}, a_{t2}, ... a_{tk}]$]. The Frequent Item Set for T is denoted by FIST, where ∀s ∈ FIST, s is the most frequent and largest set of items, which:

i. The set is not contained within another set in FIST.
ii. The set appears at least m times (we call this m number as the minimum support threshold).

FP-tree is used to find frequent item sets (also closed and maximal as well as generators) with the FP-growth algorithm (Frequent Pattern growth), which represents the transaction database as a prefix tree which is enhanced with links that organize the nodes into lists referring to the same item. The search is carried out by projecting the prefix tree, working recursively on the result, and pruning the original tree.

It allows frequent item set discovery without candidate item set generation. Two step approach:

Step 1: Build a compact data structure called the FP-tree.

Step 2: Extracts frequent item sets directly from the FP-tree [22].

D. Text Summarization

Any In today's world, due to the dramatic technological development huge amount of information is available, mostly of it is online. The World Wide Web contains billions of documents and is growing every year. The size of these documents may range from a few pages to over a thousand pages. So it is not possible to read each and every document to extract information from these documents.

Automatic summarization [9, 10, 11, 24] is a technique to create a summary by reducing a text document with a computer program. This summary preserves the most important points of the original documents. The summary of the documents helps the reader to identify key ideas and to evaluate new knowledge. Automatic summarization can be classified into two types Abstraction and Extraction [12].

Abstract Summarization: This method analyzes the document as a whole, interprets the meaning of the key ideas and creates summary. The sentences in this kind of summary may not be present in the source document.

Extraction Summarization: This is simpler than abstract one. This method works by selecting the powerful and meaningful sentences in the original text to compile a summary. The sentences in the original text that are included in the summary may change or may not change [23]. It always does not give the semantic summary. Extraction summaries are formulated by extracting key segments from the text, based on statistical analysis of individual or mixed surface level features such as word/phrase frequency, location or word to locate the sentences to be extracted [24].

III. RELATED WORKS

Recently, several research works have been carrying for text summarization. In this section some recent works are examined.

Corra, A et al. [18] have proposed a conceptual framework to promote semantic interoperability. They have considered heterogeneous human-readable documents and information as their data sources to make them meaningful and machine-readable data. They have introduced a situation to depict the working of the framework. When a user searches some keywords in search system, a huge of results is returned. From the results returned, the automatic agents do prior search and discovery. The previously built ontology meta model in e-Gov domain is applied to the pages returned. Then, from knowledge base, the system retries all the fuzzy rules defined in the context. For each ontology class, fuzzy engine calculates its membership degree and applies a label with its degree. A matching algorithm is used for matching semantic words from search string to fuzzy linguistic variables and terms. But the model has the disadvantages like though the author includes an automatic agent in the framework, developing and implementing of this agent is a challenge and defining a meta ontology model for e-Gov domain which is itself a huge knowledge domain.

Ragunath, R et al. [19] have introduced an ontology based text summarization system using concept terms. Using concepts extraction algorithm, concepts are extracted. It is used extraction summarization approaches to perform the automatic summarization

A bag of tags is created for each sentence by collecting the nodes computed by the hierarchical classifier. If a sentence is mapped to multiple sub trees in the taxonomy, all nodes are included from every sub tree.

The classifier's confidence weights are used to compute a sub tree overlap measure for each sentence. As ontology knowledge is included, it is an effective way than concept term based information retrieval methods. Stop word removal and stemming is used which removes unnecessary words.

But the model has the disadvantages as Tf-idf is used for concepts extraction, some important concepts which occur less number of times in a document may missed out.

Myaeng, Sung et al. [20] have explained about an application which takes a user query from the user which may be a keyword like cat, python and then gives summary for that keyword from the two or more web documents retrieved for that keyword. Cosine similarity is used to find two sentences from two different documents to cluster together. The size of the summary can be explicitly defined by the user. It has given statistic of the summary and original document like number of words and number of lines in a document. But the model has the disadvantages like thee summary does produce is a document with fewer sentences than the original document and It may skip some sentences which are important for the document but, may left out because of less frequency.

IV. PROPOSED WORK

The proposed model uses extractive summarization approach to perform automatic summarization. The system is consisting of two phases. In first phase (Phase 1) I have collected data sets for terrorism from the web site www.start.umd.edu/gtd [21]. I have considered this website because it is a benchmark database for world terrorism activity. The summary for all terrorism activity around the world which occurred over the years are recorded manually in this website. That's why this database can be considered as a perfect domain of worldwide terrorist activity. The datasets are used for constructing the domain knowledge. The contained of the web page are converted to text files so that further processing can be done. The proposed model consists of two phase. In first phase, text preprocessing of these data is performed to remove stop words, redundancy, duplicity as well as to reduce in size.

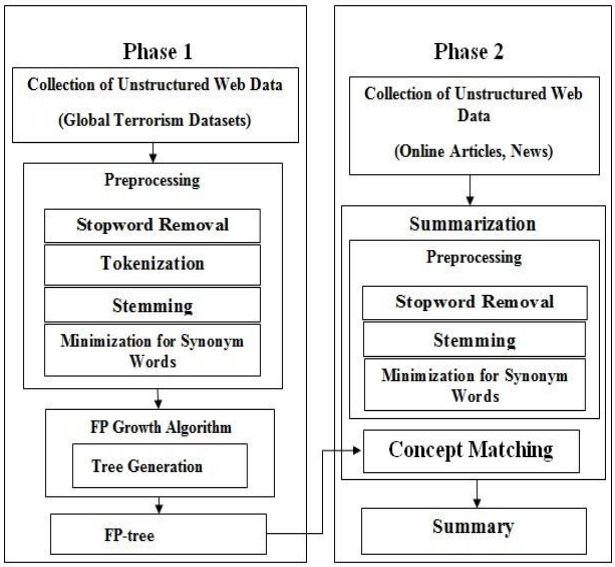

Fig. 4.1: A Block Diagram of the Proposed Method

Phase 1:- This phase consists of three components to construct domain knowledge in terms of keywords and their occurrences.

A. Collection of Unstructured Data: Different summary datasets from the Global Terrorism Database [21] are collected. Here I have considered only those data of the incident that occurred in 2014. I have designed a web crawler which collects all the related data from the mentioned website. The contained of those webpage are then converted to text file for further processing.

B. Preprocessing: Data preprocessing is an important process as it reduce the number of words to be processed by eliminating unnecessary words.

The first stage is stop word removal. This is done by comparing each word in the sentence with stop list (stop word list). Stop words are basically conjunctions, articles and prepositions present in a sentence. As stop word does not provide any semantic meaning and the number of stop words in a sentence is more, removal of stop words become necessary.

After stop word removal, the next step is stemming. Stemming is a process to obtain stem/ root of particular word in the document. It is applied to reduce different grammatical forms or word forms of a word like its noun, adjective, verb, adverb etc. For instance:

(am, are, is $\Rightarrow$ be) and (car, cars, car's, cars' $\Rightarrow$ car)

The result of this method in text document can be like: "the boy's cars are different colors" $\Rightarrow$ "the boy car be differ color".

Then the next stage is Tokenization which splits a sentence into words by following white space as the separator. Then Part-Of-Speech Tagger is applied which reads the text and assigns part of speech to each word such a noun, verb, adjective etc.

At last the synonym words are reduced from the documents. Synonyms are the words that have same meaning and can be used interchangeably. The words are exchange by most common synonyms. So, by minimizing synonyms I reduce the size of the documents. In this model I use WordNet [17] to perform all four processes.

C. FP-tree Construction: Keywords of a sentence play a vital role in semantic text summarization. I have applied FP growth algorithm to construct FP-tree for root keywords in this model. After removing the stop words as well as stemming, a weight value is assigned to each term. The weight is calculated as follows:

wt= (Frequency of the term/Total number of terms in the document)

After that ranking the individual sentence is performed as per their weight value. The weight of the sentence is found by using the following formula.

$$wt(s) = \sum_{i=1}^{n} (wt_i)/n$$

where, wt(s) is the weight of the sentence and wt_1,wt_2,...., wt_n are the weights of each terms in the sentence. Here, n is considered as total number of terms in the sentence.

Then finally, corpus which I have got from the previous stage is given as the input to the algorithm which generate FP-tree as its output based on weights of sentences which represents the domain concept with occurrence as well as related of keywords in a text document.

Phase 2:- This phase has two components for summarization of given text data.

A. Collection of Unstructured Data: The online news, articles are collected with the help of web crawler for making summary. The contained of those webpage are then converted to text file for further processing.

B. Summarization: In summarization, each paragraph of the collected article is considered for making summary. For each paragraph of the collected articles, I have again performed preprocessing tasks like removal of Stop words, stemming and Minimization of synonyms as discussed in preprocessing stage of Phase I. This preprocessing is necessary to reduce the size of paragraph and remove synonyms. I have computed similarity between each word of a preprocessed paragraph with each node in the previous constructed FP-tree. Then similarity index for processed paragraph is calculated from FP-tree. If the similarity is less than threshold value, then whole paragraph is discarded. Actual paragraph is included in the output summary when similarity measure is greater than the threshold value for the corresponding processed paragraph. In this way only those paragraphs are included which has semantic similarity with the domain knowledge. If a sentence of a paragraph is mapped to multiple sub trees in FP-tree, we include all nodes from every sub tree. I have used the following function to compute similarity measure for each sentence in a paragraph.

$$wd(p) = \sum_{sentence(i)} wt(s, i)$$

Where wd (p) is the paragraph weight with set of FP-tree keywords t and wt(s, i) is the weight value of for sentence in i^{th} position. The paragraph with the higher weights can be interpreted as the summery of a document.

Algorithm 1: (Phase 2) Text summarization for Web-based data (Sentence-wise).

Input:

1. Web based text data for which summary is required.
2. Domain Knowledge (FP-tree).

Output:

1. Summary for the original web based text data.
2. Precision.
3. Recall.

Steps:

1. Retrieval of data from the website.
2. For each sentence in the preprocessed document (Stop words free, Stemming and Minimized Synonyms),
   (a) Compare the semantic similarity measure of each sentence with each node in the FP-tree.
   (b) if semantic similarity >= threshold value , Add the original sentence to summary
   (c) otherwise discard the original sentence.
3. Calculate Precision and Recall for each original sentence.

Algorithm: (Phase 2) Text summarization for Web-based data (Paragraph-wise).

Input:

3. Web based text data for which summary is required.
4. Domain Knowledge (FP-tree).

Output:

4. Summary for the original web based text data.
5. Precision.
6. Recall.

Steps:

1. Retrieval of data from the website.
2. For each paragraph in the preprocessed document (Stop words free, Stemming and Minimized Synonyms),
   (a) Compare the semantic similarity measure of each paragraph with each node in the FP-tree.
   (b) if semantic similarity >= threshold value , Add the original paragraph to summary
   (c) otherwise discard the original paragraph.
3. Calculate Precision and Recall for each original paragraph.

V. RESULT ANALYSIS

Here in this proposed model, I have considered web data from the website Global Terrorism Database. The website contains a large datasets about the global terrorism which took place between the years 1972 to 2014. I have considered only those incidents that took place in 2014. In this period many incidents took place which is the effect of terrorism. The data for second phase is collected from many online articles and online news. I have collected those articles and news which are the sources for the website Global Terrorism Database. I have collected 30 different articles/news for this model.

In this work, I have collected online articles for making summary. I have compared our output summary with the summaries specified in the GTD database.

For evaluation of result I have measured precision and recall [23]. To compute the performance measures of the proposed model I have selected top 10 paragraphs from generated summary and have compared with the summaries provided by the GTD database. In this case I use 30 different articles.

The performance measures are defined below.

1. Precision = J/K
2. Recall = J/min(M,K)

Where, J = the number of Paragraphs that are selected correctly, K = Total Number of selected Paragraphs and M = the number of Paragraphs in summary.

Following table shows the results of the FP-tree based method.

	10	8	6	4	2	1
Precision	0.32	0.37	0.44	0.53	0.73	0.88
Recall	0.92	0.86	0.83	0.72	0.74	0.88

Table 5.1: Precision and Recall values for the Proposed System

METHOD	PRECISION	RECALL
CORRA, A ET AL. [18]	0.4348	0.4522
RAGUNATH, R ET AL. [19]	0.4824	0.4641
MYAENG, SUNG ET AL. [20]	0.4827	0.4657
PROPOSED METHOD	0.7828	0.6971

In the second experiment, I applied the proposed method for text extraction with DUC2002 document dataset. I have also compared our proposed method with some existing methods. The following table shows the better performance of the proposed method than other existing methods.

Table 4.3: Results Comparison using DUC2002 Dataset

VI. CONCLUSION

In this work, I have proposed a novel model of knowledge-based semantic analysis for text summarization of web-based dynamic text data with help of FP-tree. This model is free from ontology to find out semantic representation. I have collected data sets of the events that occurred in a month. From those datasets I have created domain knowledge with occurrence of keyword which helps in text summarization. In the future work, I have planned to design a model for summarization of multiple set of articles which will make a summary for multiple articles. A personalized semantic summarization model which will take a user query as its input and by processing many web pages returns a single summary as its output.

REFERENCES

1. Witten, I. H. (2005). Text mining. Practical handbook of Internet computing, 14-1.
2. R. Vijayarani, S., Ilamathi, M. J., & Nithya, M. Preprocessing Techniques for Text Mining-An Overview. vol, 5, 7-16
3. J. Serra, Hassel, Martin. "Evaluation of automatic text summarization." Licentiate Thesis, Stockholm, Sweden (2004): 1-75.
4. Gupta, Vishal, and Gurpreet Singh Lehal. "A survey of text summarization extractive techniques." Journal of emerging technologies in web intelligence 2, no. 3 (2010): 258-268.
5. Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177. ACM, 2004.
6. Das, Dipanjan, and André FT Martins. "A survey on automatic text summarization." Literature Survey for the Language and Statistics II course at CMU 4 (2007): 192-195.
7. Carbonaro, Antonella. Automatic Concept Extraction in Semantic Summarization Process. INTECH Open Access Publisher, 2012.
8. Ontology: http://semanticweb.org/wiki/Ontology. [Accessed in 2018].
9. Brewster, Christopher, Kieron O'Hara, Steve Fuller, Yorick Wilks, Enrico Franconi, Mark A. Musen, Jeremy Ellman, and Simon Buckingham Shum. "Knowledge representation with ontologies: the present and future." IEEE Intelligent Systems (2004): 72-81.
10. Parmeshwaran, Prerna, Juilee Rege, and Sindhu Nair. "The Use of Ontology in Semantic Search Techniques." International Journal of Computer Applications 127, no. 6 (2015): 21-24.
11. RDF Resource Description Framework: http://www.w3.org/RDF. [Accessed in 2015].
12. Decker, Stefan, Prasenjit Mitra, and Sergey Melnik. "Framework for the semantic Web: an RDF tutorial." IEEE Internet Computing 4, no. 6 (2000): 68-73.
13. Lassila, O., & Swick, R. "Resource Description Framework (RDF) model and syntax specification." W3C Recommendation 22 (1999).
14. Brickley, Dan, and Ramanathan V. Guha. "RDF vocabulary description language 1.0: RDF schema." (2004).
15. OWL Web Ontology Language overview: http://www.w3.org/TR/owl-feature. [Accessed in 2015].
16. Ma, Zongmin, Fu Zhang, Li Yan, and Jingwei Cheng. "Knowledge Representation and Reasoning in the Semantic Web." In Fuzzy Knowledge Management for the Semantic Web, pp. 1-17. Springer Berlin Heidelberg, 2014.
17. WordNet, [Online]. Available: http://wordnet.princeton.edu. [Accessed in 2015].
18. Corrêa, Andreiwid Sh, Cleverton Borba, Daniel Lins da Silva, and Pedro Corrêa. "A Fuzzy Ontology-Driven Approach to Semantic Interoperability in e-Government Big Data." International Journal of Social Science and Humanity 5, no. 2 (2015): 178.
19. Ragunath, R., Sivaranjani, N. (2015). Ontology Based Text Document Summarization System Using Concept Terms. ARPN Journal of Engineering and Applied Sciences, 10(6).
20. Myaeng, Sung Hyon, and Dong-Hyun Jang. "Development and evaluation of a statistically-based document summarization system." Advances in automatic text summarization (1999): 61-70.
21. Global Terrorism Database https://www.start.umd.edu/gtd. [Accessed in 2018].
22. Borgelt, Christian. "An Implementation of the FP-growth Algorithm." In Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations, pp. 1-5. ACM, 2005.
23. Wu, C.W, & Liu, C. L. "Ontology-based text summarization for business news articles." In Proceedings of the ISCA Eighteenth International Conference on Computers and Their Applications. 2003.
24. Das, Sufal, and Hemanta Kumar Kalita. "Semantic Model for Web-Based Big Data Using Ontology and Fuzzy Rule Mining." In Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 2, pp. 431-438. Springer International Publishing, 2016.

AUTHORS PROFILE

Dr. Sufal Das is an Assistant Professor at the Department of Information Technology (IT) in North-Eastern Hill University (NEHU), Shillong. His research interests are in the general area of Big Data Analysis, Data Mining, Machine Learning. He received his Bachelor of Technology (B.Tech.) degree in Computer Science & Engineering from West Bengal University of Technology in 2005, M.Tech. from Tezpur University in 2008 and Ph.D.(IT.) from North-Eastern Hill University in 2018. Dr. Das joined the Department of IT at Sikkim Manipal Institute of Technology (SMIT), Sikkim in August 2008. He then joined NEHU in February 2010. The author has published several research articles.

.