An Adaptive Technique for Chat Summarization

Dimple[1] Dr. Kawaljeet Singh[2] Dr. Neeraj Sharma[3]
[1]M. Phil. Student [2]Director [3]Head of Dept. & Assistant Professor
[1,2,3]Department of Computer Science
[1,2,3]University Computer Centre, Punjabi University, Patiala, India

Abstract— The chat summarization is the technique, which provides the summarization of the input chat. The chat summarization defines the words, which are used frequently in the chat. In this work, the chat summarization is generated on the basis of ontology technique. The ontology is the technique of lexical analysis in which the whole chat gets processed. In the technique of ontology words are identified from the chat. The frequency of the words is calculated on the basis of their occurrence. The frequency defines the importance of each word in the chat. The module of chat summarization is applied in python which will calculate occurrence of each word. The chat summarization module will pick the most frequent words and display it as the summarized chat. The proposed algorithm has been implemented in python and results are analyzed in terms of accuracy and execution time.

Key words: Chat Summarization, Lexical

I. INTRODUCTION

A direct human computer communication which occurs when we browse for a particular content related to a topic that happens to match with our search keyword is called information retrieval. This process involves the matching of a search keyword by a user with the documents that are related to it and contain that topic related information which is meant for a user [1]. To collect or retrieve information from bulk of data by searching algorithms is not acceptable as it takes a lot of time the concept of information extraction is utilized. Different theories or algorithms have been proposed to automate the extraction process due to the diversity and non-adaptability in terms of structured content; various constraints are usually encountered by the researchers [2]. Text summarization is a great technique that serves our purpose. In order to frame up summary; it is required to find the relevant text from the information with complete omission of unnecessary information while keeping focus on details and compile them into a document. This is not as easy as it seems to be as the common constrains of natural language processing are commonly encountered. The solution is to craft a domain independent system but there are some key points to take care of like understanding of natural language, proper representation of semantics, discourse models, and natural language generation [3]. The chances are little to get success here as domain-independent systems are bound to discover key passages along with sentence structure. Chats and forums are commonly used by students and, moreover, they offer the possibility of joint learning anytime and anywhere. It is very difficult to follow the logs of conversations and to remember what every participant said and how their contributions are linked in a coherent discourse to avoid which very few automated CSCL chat analysis systems were developed [4]. There are large datasets that include the various types of information. The information from product sales, the interaction of customer and various other sources is gathered and analyzed here. Around 15 TeraBytes of data is gathered from Facebook. The other data related to astronomy, particle accelerators is achieved from the huge telescopes that are further utilized within the biology fields [5]. In order to provide success within this research work, scalable and timely analytical processing of datasets is provided within these scenarios. There is a need of cost-effective processing which is to be derived yet. There is an introduction of Map-Reduce and Hadoop systems which are parallel data flow systems within this area [6]. In order to provide facilities to data warehousing and analytics, these methods have been utilized within the recent studies. The application of these methods can be done wither directly or with the help of high-level query language. The compilation of this system is done to a parallel dataflow graph that is further utilized to provide execution.

II. LITERATURE REVIEW

Dan Cao, et.al, (2016) reviewed in this paper [7], every one of the features that utilization metrics and idea of complex network for scoring sentences. The experiment results on single component and combinations of different features we proposed are discussed. Shortest ways demonstrated astounding for summarization, which got the highest scores for the quality of generated summary. Another contribution was the discovery of results that features combinations with a similar kind property of network indicated incredible influence to choose sentences.

Rasim Alguliyev, et.al, (2016) presented in this paper [8], a sentence scoring and selection process. The process is displayed as a multi-objective optimization issue. This paper is centered on the extractive text summarization where a summary is generated by scoring and choosing the sentences in the source text. At first it assesses the score of each sentence and afterward chooses the most representative sentences from the text by considering that semantic similarity between chose sentences will be low. The proposed show endeavors to find balance amongst coverage and redundancy in a summary. For taking care of the optimization issue a human learning optimization algorithm is used.

Narendra Andhale, et.al, (2016) presented is this paper [9], the comprehensive survey of both the approaches in text summarization. This survey paper covers extractive and abstractive summarization techniques. Summarization system should produce an effective summary in a brief span with less redundancy having grammatically correct sentences. Both extractive and abstractive technique yields good result as indicated by the context in which they utilized. The surveyed literature opens up the testing area for hybridization of these methods to produce informative, all around compressed and readable summaries.

Rupal Bhargava et.al, (2017) proposed in this paper [10], a strategy utilizing which one can break down various languages to find sentiments in them and perform sentiment

analysis. The strategy leverages diverse techniques of machine learning to dissect the text. Machine translation is utilized as a part of the system to give the component of dealing with various languages. So the system proposed utilizes text summarization process to extract important parts of text and after that utilizations it to examine the sentiments about the specific subject and its aspects. Experiment demonstrates that proposed strategy can deliver promising results.

Archana N.Gulati, et.al, (2017) discussed in this paper [11], that text summary is a reduction of original text to condensed text by choosing what is important in the source. Text summarization is required when individuals need a gist of a specific topic from at least one sources of information accessible on the web. A novel procedure for multi document, extractive text summarization is proposed. The system accomplishes an average precision of 73% over multiple Hindi documents. The summary generated by the system is discovered near summary generated by humans. The Precision, Recall and F-score values demonstrates good accuracy of summary generated by the system.

Manisha Gupta, et.al, (2016) presented in this paper [12], a novel approach for text summarization of Hindi text document based on some linguistic principles. Proposed system is tested on different Hindi sources of info and accuracy of the system in type of number of lines extracted from original text containing important information of the original text document. Info text size can be decreased to 60% - 70 % with the assistance of proposed system. System generates the extractive summary given by the client i.e. it doesn't generate the summary of the text on the premise of the semantics of the text.

III. RESEARCH METHODOLOGY

The research work focuses on the chat summarization of the IRC chat from the computer center of Punjabi University, Patiala. The summarization technique includes following steps:

1) Input Data: In the first step, the data is given as input and input data is the chat data which can either be in the excel sheet or the real time data which is extracted using the chat API's.
2) Pre- processing: In the pre-processing phase, the data is pre-processed. It is tokenized and stop words are removed from the data.
3) Chat Summarization: In the last step, the rating to each word is given on the basis of their occurrences and the words with maximum rating is considered as most important words that are included in the final chat summary and others are removed.

The pattern based algorithm is the algorithm which generates patterns of the input data. The weight is assigned to each word, character in the chat for generation of final chat summary.

A. Proposed Algorithm

– Input: IRC chat
– Output: Summary

begin
Process input data to analyze their occurrences
Get size(s) of the file
for i:1 to s do
assign rating to each line according to occurrence
extract repeated text lines for summary
end for
compute size (u) of unrepeated lines
for i:1 to u do
assign rating to each word and alphabet
extract high rated words and alphabets for summary
end for
Display high rated lines, words and strings as output summary
End

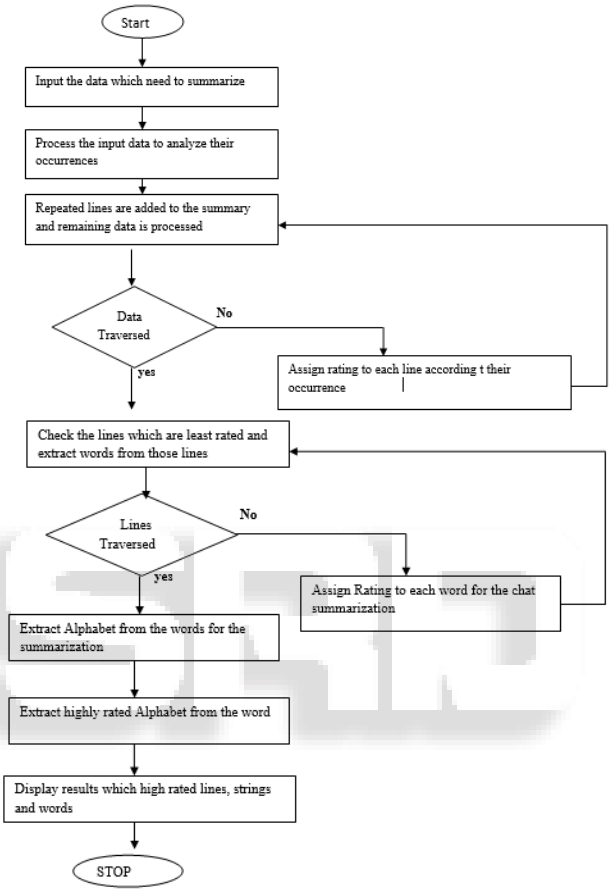

Fig. 1: Flowchart of Proposed Chat Summarization Technique

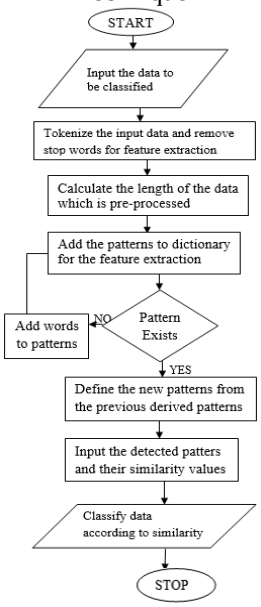

Fig. 2: Flowchart of Sentiment Analysis

IV. EXPERIMENTAL RESULTS

The proposed work has been implemented in Python and the achieved results have been compared with the existing approach on the basis of the performance parameters which are accuracy and execution time.

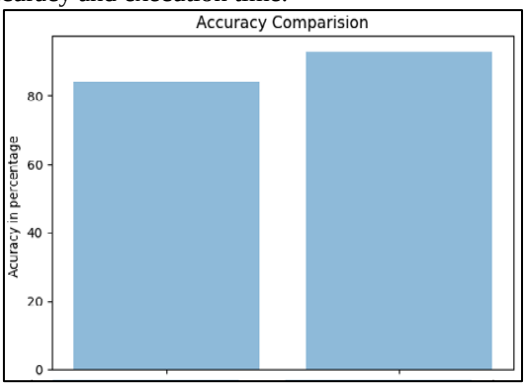

Fig. 3: Accuracy Comparison

As shown in figure 3, the accuracy of pattern based algorithm and E-patterns based algorithm is compared in terms of accuracy and it is been analyzed that accuracy of enhanced algorithm is more due to batter analysis of the data.

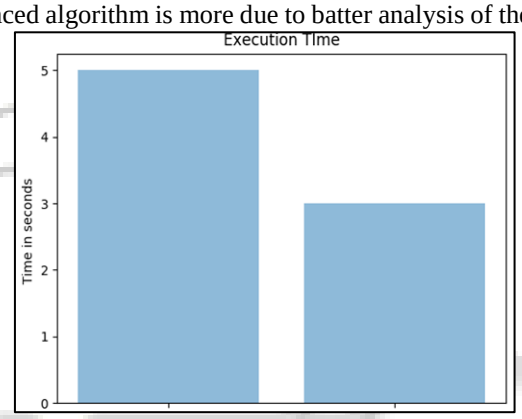

Fig. 4: Execution time

As shown in figure 4, the execution time of proposed and existing algorithm is terms of execution time. It is been analyzed that enhanced pattern based algorithm is less execution time.

V. CONCLUSION

In this work, it has been concluded that chat summarization is technique which generate the summary of the chat. The efficient novel technique is proposed in this paper which generate summary of the chat using string, word and character occurrence. The proposed algorithm is implemented in python and compared with the existing algorithm of chat summarization. The results are generated in python in the terms of accuracy and execution time.

REFERENCES

[1] D. Shen, J.-T. Sun, H. Li, Q. Yang, and Z. Chen, 2007,"Document summarization using conditional random fields", IJCAI, vol. 7, ,pp. 2862-2867

[2] Y. Gong and X. Liu, 2001 "Generic text summarization using relevance measure and latent semantic analysis," in Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, , pp. 19-25

[3] R. Mihalcea, 2005, "Language independent extractive summarization," in Proceedings of the ACL 2005 on Interactive poster and demonstration sessions. Association for Computational Linguistics, pp. 49-52

[4] N. Lalithamani, R. Sukumaran, K. Alagamrnai, K. K. Sowmya, V. Divyalakshmi, and S. Shanmugapriya, 2014,"A mixed-initiative approach for summarizing discussions coupled with sentimental analysis," in Proceedings of the 2014 International Conference on Interdisciplinary Advances in Applied Computing. ACM, p. 5

[5] Prachi Shah, Nikita P. Desai, 2016, "A Survey of Automatic Text Summarization Techniques for Indian and Foreign Languages" International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)

[6] Jyoti Yadav, Dr. Yogesh Kumar Meena, 2016, "Use of Fuzzy Logic and WordNet for Improving Performance of Extractive Automatic Text Summarization", Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI)

[7] Dan Cao, Liutong Xu, 2016, "Analysis of Complex Network Methods for Extractive Automatic Text Summarization", 2nd IEEE International Conference on Computer and Communications

[8] Rasim Alguliyev, Ramiz Aliguliyev, Nijat Isazade, 2016, "A Sentence Selection Model and HLO Algorithm for Extractive Text Summarization", IEEE

[9] Narendra Andhale, L.A. Bewoor, 2016, "An Overview of Text Summarization Techniques", IEEE

[10] Rupal Bhargava and Yashvardhan Sharma, 2017, "MSATS: Multilingual Sentiment Analysis via Text Summarization", IEEE

[11] Archana N.Gulati, Dr.S.D.Sawarkar, 2017, "A novel technique for multi-document Hindi text summarization", International Conference on Nascent Technologies in the Engineering Field (ICNTE-2017)

[12] Manisha Gupta, Dr.Naresh Kumar Garg, 2016, "Text Summarization of Hindi Documents using Rule Based approach", International Conference on Micro-Electronics and Telecommunication Engineering

