International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 03 | Mar 2020 www.irjet.net p-ISSN: 2395-0072

PDF Extraction Using Data Mining Techniques

Madhuri Badhe[1], Vrushali Thakur[2], Pooja Patil[3], Rukhsar khan[4], Prof. N. L. Bhale[5]

[1,2,3,4]Student, Dept. of Information Technology, Matoshri College of Engineering and Research Centre, Maharashtra, India
[5]Head of Department, Dept. of Information Technology, Matoshri College of Engineering and Research Centre, Maharashtra, India

---

Abstract - *In this new era, where tremendous information is available on the internet, it is most important to provide the improved mechanism to extract the information quickly and most efficiently. It is very difficult for human beings to manually extract the summary of a large documents of text. There are plenty of text material available on the internet. So there is a problem of searching for relevant documents from the number of documents available, and absorbing relevant information from it. In order to solve the above two problems, the automatic text summarization is very much necessary. Text summarization is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings.*

Keywords: *Documentations, Text, Summarization, Quickly, easy*

1. INTRODUCTION

Summary can be de_ned as a brief and accurate way of representing the important concepts of the given source documents. Humans, during the process of text summarization, understand the concept of source document and create a summary which conveys the essence of the document whereas in automated systems this is a complex task. As the quantity of information available in electronic format continues to grow, research into automatic text summarization has taken huge importance. There are two types of summary Extractive and Abstractive. Abstractive summary represents use of . (NLP) whereas Extractive summary is based on copying exact sentences from source document. Presently it is not possible that the computer can understand every aspect behind Natural Language processing. So, our Scope is limited to Extractive based summary.

1.1 Aim

Our aim is to identifies the most important points of a text and expresses them in a shorter document. Summarization process:

1. interpret the text;
2. extract the relevant information (topics of the source);
3. condense extracted information and create summary representation;
4. Present summary representation to reader in natural language.

1.2 Motivation of the Project

As we can see in our daily lives , people not to take more interest to read the books and big documents because it takes more time and very boring mostly for the students so that we can introduce such a system that which can reduce the text in such a file, pdf or any and give output only in summarized form so anyone one can understand the thing in this document easily.

1.3 Objectives

Automatic summarization involves reduces a text le into a passage or paragraph that conveys the main meaning of the text. The searching of important information from a large text file is very difficult job for the users thus to automatic extract the important information or summary of the text file.

This summary helps the users to reduce time instead Of reading the whole text file and it provide quick Information from the large document. In today's world to extract information from the World Wide Web is very easy. This extracted information is a huge text repository.

With the rapid growth of the World Wide Web (internet), information overload is becoming a problem for an increasing large number of people. Automatic summarization can be an indispensable solution to reduce the information overload problem on the web.

2. LITERATURE SURVEY

Paper 1: CyberPDF: Smart and Secure Coordinate-based Automated Health PDF Data Batch Extraction

Data extraction from files is a prevalent activity in today's electronic health record systems which can be laborious.

Paper 2: PDF Scrutinizer: Detecting JavaScript-based attacks in PDF documents

For a long time PDF documents have arrived in the everyday life of the average computer user, corporate businesses and critical structures

Paper 3: Data Mining Based Strategy for Detecting Malicious PDF Files

Portable Document Format (PDF) is one of the widely-accepted document format. The file can be viewed on any information processing system with a PDF viewer in the year 2014.

Paper 4: A new method of information extraction from PDF files

With the rapid increase of the PDF files in Internet, how to manage and search PDF files efficiently and quickly has become an urgent problem to be solved.

3. ARCHITECTURE

3.1 Problem Statement / Definition

This project describes a system for the summarization of single and multiple documents. The system produces multi as well as single document summaries using data mining techniques for identifying common terms across the set of documents. For each term, the system identifies representative passages that are included in the final summary. Results of our evaluation are also presented.

3.2 Proposed Architecture

We are introducing a system system that allows user to extract meaningful information from a particular pdf, by using text characteristic algorithm . User has to upload the file into our system and system will get process on that file and give output to user.

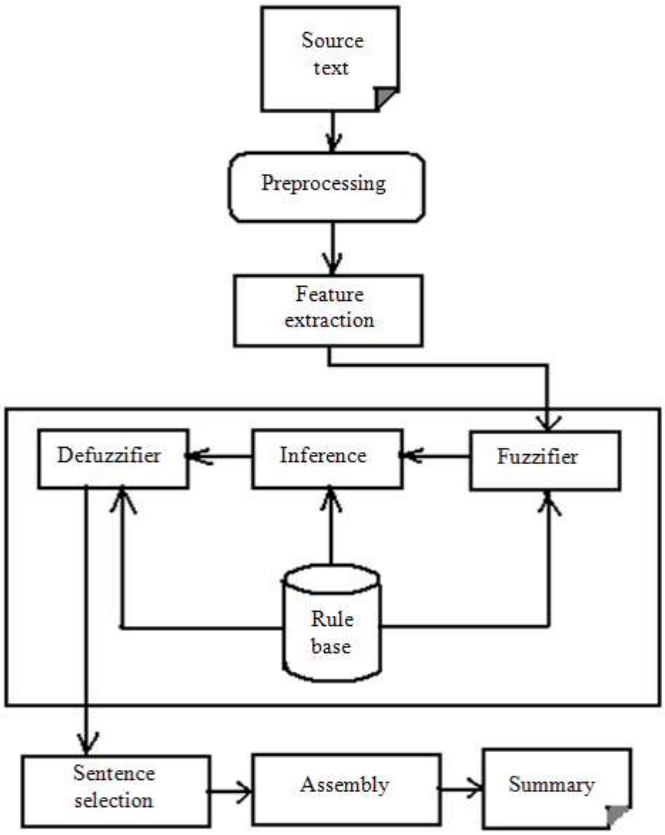

Figure 3.1: Proposed Architecture Diagram

CONCLUSION

The Conclusion of this project is that the client will get an Web application that will execute on client side and get the summary of the input document as per his/her requirement. The effective diversity based method combined with K-mean Clustering algorithm to generating summary of the document. The clustering algorithm is used as helping factor with the method for finding the most distinct ideas in the text. The results of the method supports that employing of multiple factors can help to find the diversity in the text because the isolation of all similar sentences in one group can solve a part of the redundancy problem among the document sentences and the other part of that problem is solved by the diversity based method.

REFERENCES

[1] Adobe Systems Incorporated PDF Reference 6th edn, November 2006.
[2] S.Y. Bai, "Method Research and System Design of Printed Mathematical Formula Recognition Based on SVM", Shen Yang: ShenYang University of Technology, pp. 1-66, 2015.

[3] L.Y. Lin, L.C. Gao, Z. Tang, "Research on Mathematical Formula Identification in Digital Chinese Documents", Act a Scientiarum Naturalium Universitat is Pekinens is, vol. 50, pp. 17-24, January 2014.
[4] D.R. Li, T.D. Xu, "Research on an Extraction Method for Mathematical Formulas Embedded in Printed Documents", Computer Applications and Software, vol. 31, pp. 102-105, April 2014.
[5] Z.F. Guo, "Mathematical Formula Feature Extraction and Locating in Chinese Scanned Printed Document", GuangXi: GuangXi Normal University, pp. 1-35, 2010
[6] Y.S. Guo, N.T. Tan, L. Hang, C.P. Liu, "An Identification Method for Mathematical Expressions in Scanned Chinese Document", Journal of Chinese Information Processing, vol. 22, pp. 83-87, July 2008
[7] X.D. Tian, N. Hao, "Mathematical Formula Extraction Method from Printed Document Based on Fuzzy Classification", Computer Applications, vol. 27, pp. 2036-2038, August 2007.
[8] Z.W. Zhang, F.R. Kong, W.L. Liu, Q. Long, Y.B. Liu, "Extraction of Mathematical Expressions in Printed Chinese Technical Documents", Journal of Chinese Information Processing, vol. 21, pp. 86-91, July 2007.
[9] J.M. Jin, X.H. Han, Q.R. Wang, "Mathematical Formulas Extraction", Proceedings of the Seventh International Conference on Document Analysis and Recognition, vol. 2, pp. 1138-1141, 2003