Survey on Data Extraction using Text Mining

Akshitha Shetty P[1], Sumana[2], Sushmitha K Shetty[3], Ujwala B M[4], Sadhana.B[5]
[1,2,3,4] Student, Information Science and Engineering, Canara Engineering college Benjanapadavu,
Associate Professor, Department of ISE, Canara Engineering College, Benjanapadavu India

***Abstract*— This paper is focused on data extraction. Thses days users try to store their data in computer rather than storing it manually for security purpose and for better efficiency. As a result, large amount of data stored in document format within computers. These documents can be of any form i.e, structured, semi structured and unstructured format. Retrieving useful data from these large amount of data is a very tedious work .In this scenario text mining is a useful technic. Discover knowledge from unstructured text is major problem but text mining technic is one of the inspiring research area to clear all these kind problems. This paper gives an brief description about concept of application, issues and the tools used for text mining. In the last decades explosion of information and communication technologies has led to a whole new scenario concerning people's accessibility to new job opportunities and company's options for employing the right person for the right job. In this work, we present a set of techniques that makes the whole recruitment process more effective. We have implemented a system that models the candidate's resume in a Structured Data. Finally, it presents the results to the recruiter based on their criteria. In this case today's technology is used which basically helps the recruiter to get only the required resumes from the huge slot. The aim is to obtain a resume from bulk of the resume pool and to extract the min data on particular user defined condition.**

This paper represents an implementation model turned Data. Finally, it presents the results to the recruiter based on their criteria. In this case today's technology is used which basically helps the recruiter to get only the required resumes from the huge slot. The aim is to obtain a CVs from bulk of the CV pool and to extract the min data on particular user defined condition.

Keywords :Text Mining, Information Extraction, Knowledge Discovery from Databases

I.INTRODUCTION

As we know these days, there is a huge Imbalance in recruitment of employees and the number of candidates that apply. There are many number of CVs received each day for vacancy. Thus it is a highly tedious job to sort these out without missing on a prospective candidate. By checking manually there is a chance of missing some of the CVs. For example, consider naukri.com website in which many people will be applying for various job description. So in this paper Text Mining Technique is implemented. Text mining is referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high quality information from text. Text mining usually involves the process of structuring the input text, deriving patterns within the structured data, and finally evolution and interpretation of the output. In on-line recruitment systems, candidates typically upload their CVs in the form of a document with a loose structure, which must be considered by an expert recruiter. This incorporates a great asymmetry of resources required from candidates and recruiters, resulting in candidates uploading the same CV in numerous HR agencies that become overwhelmed with thousands of CVs. In this work, we follow a different approach in the CV submission process.

In the proposed system, we mandate that applicants submit their CVs in a structured way. This examine the candidates professional qualifications and his personality. The forms designed are divided in the 4 sections. In the first section, which is the education and qualification section, the candidate fills in his academic degrees (BSc, MSc, PhD) and professional qualifications. The candidate is expected to be able to prove all entered information in this section. In the second section, the experience section, there are questions about the applicant's professional history. These include his years of experience, the candidate's loyalty, his former position titles and the organizational culture of his previous jobs. In the third section, the specialization section, in which the candidate is asked about their respective branches in which they have completed their Qualification.

II.LITERATURE SURVEY

In order to understand the development of research in text mining in the last years, it is important to conduct a literature review to understand the different fields of application through which text mining has evolved as well as to identify research gap. Therefore in the next sections the research process, methodology and findings of the literature review are presented.

[1]Raymond J. Mooney and Un Yong Nahm, suggested a new framework for text mining based on the integration of *Information Extraction* (IE) and Knowledge Discovery from Databases (KDD), a.k.a. *data mining*. KDD and IE are both topics of significant recent interest. KDD considers the application of statistical and machine-learning methods to discover novel relationships in large relational databases. IE concerns locating specific pieces of data in natural-language documents, thereby extracting structured information from free text. However, there has been little if any research exploring the interaction between these two important areas. In this paper, we explore the mutual benefit that the integration of IE and KDD for text mining can provide

[2]Evanthia Faliagka, Konstantinos Ramantas, Athanasios Tsakalidis, Manolis Viennas, discovered that rapid development of modern Information and Communication technologies (ICTs) in the past few years and their introduction into people's daily lives has led to new circumstances at all levels of their social environment(work,interpersonalrelations,entertainment, etc).

People have been steadily turning to the web for job seeking and career development, using web 2.0 services like LinkedIn and job search sites (Bizer, 2005). On the other hand, a lot of companies use online knowledge management systems to hire employees, exploiting the advantages of the World Wide Web. These are termed e-recruitment systems and automate the process of publishing positions and receiving CVs.

[3]Shaidah Jusoh and Hejab M. Alfawareh, said that, In this modern culture, text is the most common vehicle for the formal exchange of information. Although extracting useful information from texts is not an easy task, it is a need of this modern life to have a business intelligent tool which is able to extract useful information as quick as possible and at a low cost. Text mining is a new and exciting research area that tries to take the challenge and produce the intelligence tool. The tool is a text mining system which has the capability to analyze large quantities of natural language text and detects lexical and linguistic usage patterns in an attempt to extract meaningful and useful information [1]. The aim of text mining tools is to be able to answer sophisticated questions and perform text searches with an element of intelligence

[4]According to K.L.Sumathy, M.Chidambaram, said that, Nowadays there is an increasing trend in the usage of computers for storing documents. As a result of it substantial volume of data is stored in the computers in the form of documents. The documents can be of any form such as structured documents, semi-structured documents and unstructured documents. Retrieving useful information from huge volume of documents is very tedious task. Text mining is an inspiring research area as it tries to discover knowledge from unstructured text.

[5]Sonali Vijay Gaikwad, Archana Chaugule, Pramod Patil, said that, Nowadays most of the information in business, industry, government and other institutions is stored in text form into database and this text database contains semi structured data. A document may contain some largely unstructured text components like abstract additionally few structured fields as title, name of authors, date of publication, category, and so on. Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. The great deal of studies done on the modeling and implementation of semi structured data in recent database research. On the basis of these researches information retrieval techniques such as text indexing methods have been developed to handle unstructured documents. In traditional search the user is typically look for already known terms and has been written by someone else. The problem is in result as it is not relevant to users need. This is the goal of text mining to discover unknown information which is not known and yet not written down.

[6]Samiddha Mukherjee, Ravi Shaw, Nilanjan Halder, Satyasaran Changdar, said that, In layman terms Data-mining can be related to human cognitive mind where based on previous knowledge and experience we can relate things happening around us or sometimes even predict the future. Data mining is a process of searching data from a pool of data like database, web-servers, cloud based servers etc. and provide a pattern or relationships among those data to produce desired information. This paper conducts a formal review of the concept of data-mining, the standard tasks involve in data-mining, its applications in day to day field, techniques and methodology.

[7]According to author Ramzan Talib_, Muhammad Kashif Hanify, Rapid progress in digital data acquisition techniques have led to huge volume of data. More than 80 percent of today's data is composed of unstructured or semi-structured data. The discovery of appropriate patterns and trends to analyse the text documents from massive volume of data is a big issue. Text mining is a process of extracting interesting and nontrivial patterns from huge amount of text documents. There exist different techniques and tools to mine the text and discover valuable information for future prediction and decision making process. The selection of right and appropriate text mining technique helps to enhance the speed and decreases the time and effort required to extract valuable information. This paper briefly discuss and analyze the text mining techniques and their applications in diverse fields of life. Moreover, the issues in the field of text mining that affect the accuracy and relevance of results are identified.

III.METHODOLOGY

In the contemporary world the text is the most common means for exchanging information. The data stored in the computer can be in any one of the form (i) structured (ii) semi structured and (iii) unstructured. The data stored in databases is an example for structured datasets. The examples for semi structured and unstructured data sets include emails, full text documents and HTML files etc. Huge amount of data today are stored in text databases and not in structured databases. Text Mining is defined as the process of discovering hidden, useful and interesting pattern from unstructured text documents. Text Mining is also known as Intelligent Text Analysis or Knowledge Discovery in Text or Text Data Mining. Approximately 80% percent of the corporate data is in unstructured format. The information retrieval from unstructured text is very complex as it contains massive information which requires specific processing methods and algorithms to extract useful patterns. As the most likely form of storing information is text, text mining is considered to have a high value than that of data mining. Text mining is an interdisciplinary field which incorporates data mining, web mining, information retrieval, information extraction.

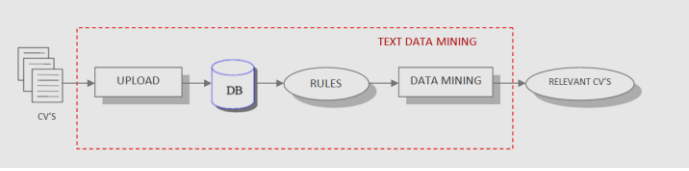

Figure1: Architecture of Text Data Mining

The CVs are sent by the candidates to the companies and these CVs are uploaded to company's Database. By following

some specific requirements for the company CVs are filtered accordingly using text mining. Then the CVs are segregated using data mining and the relevant CVs are accessed by the admin as shown in Figure1.

IV.CONCLUSION

The conclusion of this paper have described the implementation of usage of Text Mining. This has reduce the execution time of a particular task and hence has been helpful in accomplishment of the objective. Here by implementing this paper, which are able to avoid the manual work & increase the efficiency of the recruitment process.

REFERENCES

[1] Ramzan Talib, Muhammad Kashif Hanify, "*Text Mining: Techniques, Applications and Issues*", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7 No. 11, 2016.

[2] Samiddha Mukherjee, Ravi Shaw, Nilanjan Halder, Satyasaran Changdar, "*A Survey Of Data Mining Applications And Techniques*", International Journal of Computer Applications, Volume 6(5), 2015.

[3] Sonali Vijay Gaikwad, Archana Chaugule, Pramod Patil, *"Text Mining Methods And Techniques*", International Journal of Computer Applications, Volume.85, No.17, January 2014.

[4] K.L.Sumathy, M.Chidambaram, "*Text Mining: Concepts, Applications, Tools and Issues – An Overview*", International Journal of Computer Applications (0975 – 8887) Volume 80 – No.4, October 2013.

[5] Shaidah Jusoh, Hejab M. Alfawareh, "*Techniques, Applications And Change in Issue In Text Mining*", International Journal of Computer Applications Volume.9, Issue.6 N0.2, November 2012.

[6] Evanthia Faliagka, Konstantinos Ramantas, Athanasios Tsakalidis, Manolis Viennas, *" An Integrated E-Recruitment System form CV Ranking based on Asp*"**,** WEBIST 2011 - 7th International Conference on Web Information Systems and Technologies

[7] "*Text Mining with Information Extraction*", Multilingualism and Electronic Language Management: Proceedings of the 4th International MIDP Colloquium, September 2003, Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.) pp.141-160, Van Schaik Pub., South Africa, 2005.