I.J. Information Technology and Computer Science, **2018, 9, 21-31**
Published Online September 2018 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2018.09.03

A CV Parser Model using Entity Extraction Process and Big Data Tools

Papiya Das
KIIT University, School Of Computer Engineering, Odisha, Bhubaneswar 751024
E-mail: papiyanita895@gmail.com

Manjusha Pandey and Siddharth Swarup Rautaray
KIIT University, School Of Computer Engineering, Odisha, Bhubaneswar 751024
E-mail: manjushafcs@kiit.ac.in, siddharthfcs@kiit.ac.in

Received: 08 February 2018; Accepted: 14 July 2018; Published: 08 September 2018

Abstract—Private organizations like offices, libraries, hospi-tals make use of computers for computerized database, when computers became a most cost-effective device.After than E.F Codd introduced relational database model i.e conventional database. Conventional database can be enhanced to temporal database. Conventional or traditional databases are structured in nature. But always we dont have the pre-organized data. We have to deal with different types of data. That data is huge and in large amount i.e Big data. Big data mostly emphasized into internal data sources like transaction, log data, emails etc. From these sources high-enriched information is extracted by the means of process text data mining or text analytics. Entity Extraction is a part of Text Analysis. An entity can be anything like people, companies, places, money, any links, phone number etc. Text documents, bLogposts or any long articles contain large number of entities in many forms. Extracting those entities to gain the valuable information is the main target. Extraction of entities is possible in natural language processing(NLP) with R language. In this research work we will briefly discuss about text analysis process and how to extract entities with different big data tools.

Index Terms—Computerized Database, Conventional Database, Entity Extraction, Natural Language Process-ing(NLP), Temporal Database, Text data mining, Text analytics.

I. Introduction

Big data are complex data sets with traditional [1]data processing application [2].Big data makes use of predic-tive analysis, user behavior analysis and some value extrac-tion method. Data sets are growing rapidly and are gath-ered by numerous sensors. Big Data deal with challenges like capture, storage, analysis [3] etc. We can computation-ally analyze data sets [4] to reveal patterns, trends and associations. Multidimensional big data represented as ten-sors. However tensors can be efficiently handled by tensor based computation. Among these major is text data. Text data are already defined in a predefined data format. We have to ana-lyze the data using text analytics method. It is about deriving of high quality structured data from unstructured text. It enriches customer master data to produce new consumer and deter-mines the where about of products and services. Actually text databases are huge collection of documents. Database collect information from books, articles etc. Retrieval of information is done from the text based documents [5].This paper is about CV parsing, the recent key part of text analytics process. It depends upon entity extraction. It is parsing and extraction of entities from raw text. It can be a searching process for any integrated software system. Candidate can use it mobile friendly.CV parsing is generally conversion of free form CV document into XML format structured form of informa-tion. Generally, CV parsing tools are used by recruitment com-panies for storing and evaluation of data. Hadoop can be used sometimes as a big data tool for many other applications like for the word processing process Hadoop MapReduce used for reducing the number of consecutive words in a system [6].Text analysis is parsing of texts to extract facts which is readable by machine from set of unstructured archives. Text analysis and text analytics both are very different from each other. Analysis depicts the process of analyzing of texts computationally. But the analytics is for gaining patterns from text i.e the process of text mining [7].Facts extracted through text analysis from long articles or large text bodies [8] are stored in database for further analysis. These facts stored in databases can be analyzed for gaining natural languages summary. It also has applications in information retrieval process. Big Data plays a vital role in Entity Extraction process. Suppose from a large document Named Entities like People, places and products and values like emails, links, telephone numbers have to be extracted as in fig number 1. So, all these entities contain the informations on a piece of text. So, for extracting these entities from that piece of text, firstly we should identify the named entities and values & next we extract them [9].So, it applies the concept of knowledge disc overy process of big data.

Further section 2 & 3 describes Entity extraction process, Section 4 describes previous years work, Section 5 & 6 is all about proposed model, Section 7 concludes the paper.

II. Process of Entity Extraction

Entity extraction also called Entity Name Extraction or named entity recognition(NER). NER is subtask of information extraction process to identify the named entities in text into different types of facts like name of a per-son, organizations, values etc. It helps in transformation of un-structured text to structured text [10].

Fig 2 shows how entity extraction distinguishes the named entities from the text. Input can be in any format, such as documents, spreadsheets, webposts etc. In these formats text is in un-structured format. Entity extraction can identify entities i.e people, places etc; numerical expressions i.e date, time, phone numbers etc and also temporal expressions like frequency, durations etc. As example: data analyst, journalists have millions of documents for study & review. At the first stage they might not know what the information contains actually. So, entity extraction shows an useful view of unstructured unknown data sets by extracting information [11]. And then stores the converted structured data i.e named entities and phone numbers also in a corpus for further analysis.

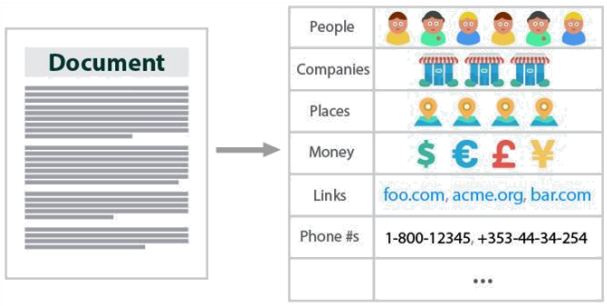

Fig.1. Entity Extraction Process Example

Fig.2. Extraction of named Entity

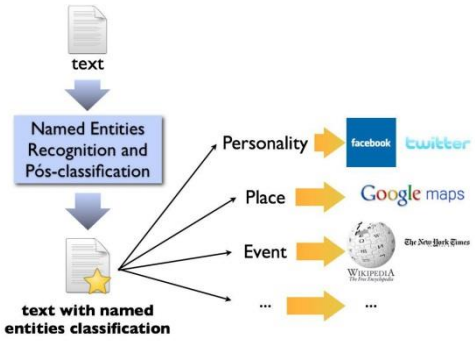

Fig.3. Process of Entity Extraction

III. Working of Entity Extraction

Extraction rules are based on Semantic technologies. These technologies sometimes have to face some language issues for correctly identifying the entities. It will be easy for human being to distinguish names, places, organizations etc. But in case of machines, it become more complex because of ambiguous nature of grammar. Keywords based system are unable to differentiate between every possible meanings of a word. As example ”Orange” can be a name of a fruit, any color or name of any country. Semantic technologies of understanding context [12] in text includes:

Entity Relation Extraction: This connects relationships shared between different entities. It also reveals any connec-tions or events shared among these entities. The main aspect is it also establishes the indirect relationships amongst indirect connections.

Linking: It identifies links of any entities in a cor-pus. Example: linking identify any place in the corpus and link that place on the map.

Fact Extraction: It helps in extracting all informations asso-ciated with that entity.

Fig.4. Working of Entity Extraction

IV. Related Works

The research work carried in the area of big data with varied domains like banking, healthcare, education, insurance, text analytics etc. The problem associated with

huge amount of data could not be processed by structured query language based queries used in relational database management system be used for analysis of big data. So text mining or text analytics is used. Text mining is the process of distilling actionable insights from text. When a satellite is taking image for social media pictures & traffic information system. It is too much information to handle and to organize. These are the bunch of text which is nearly impossible to organize quickly. Text mining is all about organizing the unorganized text. The data scientists of IBM splits big data into 4 Vs i.e volume, variety, velocity, veracity. Sometimes a fifth V i.e value [6] also considered. The objective of this division is to extract business value and to improve their organizational power. As business value is more focused in an organization. VOLUME the word itself is describing about its nature. As the world record, every year data generation graph is increasing. So it merely thinks about small storage capacities. Datas increasing nature starts from bits to bytes, bytes to kilobytes and still continuing with the increase of data as in Table 1. This generates more complexity within the data and more categorization will be done in future to handle the data. VARIETY deals with both structured and unstructured data [13].Traditional structured data includes users bank account statements. Audio images, Video images, logs are structured data mounted by unstructured data. VERACITY is the biggest hurdle in big data. Its main aim is to maintain the cleanliness of data. And remove the inconsistent data. VELOCITY [14] is the rate at which data is generating day by day. Fig number 5 depicts the 4 V's. It is a tremendous and extensive method.

Fig.5. 4 V's of Big Data

Table 1. Data Measurement Charts With Its Units And Values

Data measuring units	Values
Bit	0,1
Byte	8 bits
Kilobyte	1000^1bytes
Megabyte	1000^2bytes
Gigabyte	1000^3bytes
Terabyte	1000^4bytes
Petabyte	1000^5bytes
Exabyte	1000^6bytes
Zettabyte	1000^7bytes
Yottabyte	1000^8bytes

Fig.6. Text mining Process

1) Text Analytics Techniques:

Text Analytics is also a [15] predictive analysis method. When the training data sets or texts comes, user categorized the texts into different portions or texts for classification. This classification process includes the text preprocessing steps. Steps are cleaning, tokenization, POS tagging, transformation, evaluation, Selection of attributes, insights identification etc. When the raw text or raw data comes, statistical or linguistic techniques applied for the text analysis. Then the texts according to their taxonomical behavior is categorized. Then the concepts and patterns is extracted for getting relationships in large amounts of text. After that accuracy level of the model is being checked. When there is unstructured, ambiguous data which is difficult to process, text preprocessing method is used as shown in fig 5.

Preprocessing includes different steps as in fig 6.First step is text cleaning which involves removal of unwanted or inconsistent data. Example: ads in webpages, popups coming in websites etc. Tokenization is breaking of textual contents into words, symbols named as tokens. Part of Speech Tagging includes transformation, selection, mining, evaluation. After the tokenization process tagging assigned to each token. Text transformation is also attribute generation. Its main approaches are Bag of words and Vector Space. Feature selection is also Attribute Selection. Main aspect of this method is to remove redundant and irrelevant features.

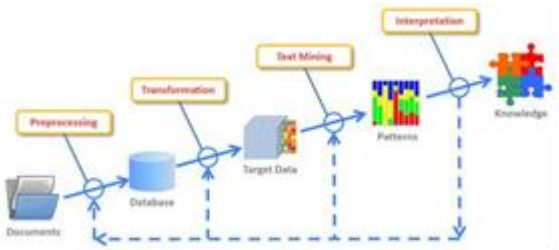

Fig.7. Text preprocessing method

Text analytics techniques are based on different applications of Text analytics as in fig 7.Sentiment analysis techniques, text blogs in python or Natural language Toolkits. Many top free text analytics. Software are there for analysis of text. Topic modelling and text mining are the most wanted platforms for evaluation of data and getting targeted data set. Text analytics is also known as Text Mining. Mining are of different types. Data Mining, Text Mining, Web Mining etc. Data mining is extraction of valuable meaningful information from huge amount of data collected from data warehouse. Text Mining is nothing other than text analytics. We have discussed about text analytics in the introduction part. Information gathered by web mining is the collection of data from traditional data mining methodologies and information covered over WORLD WIDE WEB (WWW).Other mining techniques are Sequence Mining, Graph Mining, Temporal Data Mining, Spatial Data Mining(SDM), Distributed Data Mining(DDM) and Multimedia Mining. Different applications of text mining are Sentiment Analysis, Feedback Analysis, Competitive & computing intelligence, National security issues, Accuracy and precision measurement of data, Monitoring of social media, Management of e-record and data.

2) Applications of Text Analytics:

i) Banking and Securities: Applications of Big data in banking is very crucial. The SEC (Securities Exchange Commission) use big data for monitoring stock and financial market [16].
ii) Communications, Media and Entertainment: By this we collect and an analyze consumer insights. It tries to understand patterns of real time contents of data.
iii) Health care Providers: Hospitals uses data from mobile [17] for millions of patients for detecting several lab tests.
iv) Education: Big universities and institutions uses big data tools for managing student information according to a specific key or primary key i.e roll number. Tools may be SQL method.
v) Manufacturing and Natural Resources: In this approach big data use predictive modelling for integrating large amounts of data. The types of data may vary like it can be graphical data [18], text data or may be temporal data.
vi) Government: Government uses big data in various fields like in SSA(Social Security Administration) & FDA(The Food and Drug Administration).
vii) Insurance: In the insurance industry they predicts the customers insight and predictive behaviour from different social media, GPS enabled device, CCTV footage [19] etc.
viii) Finance: Big Data uses technical Analysis in the Financial market.
ix) Internet of Things: Big data and IOT works concur-rently. The targetted data is extracted from the IoT device for preprocessing and it provides mapping of device interconnectivity.
x) Sports: Big data uses prediction modelling in sport sensors for performance improvement of players. It also can predict winners in a match using Big Data Analytics.

Fig.8. Text Analytics Applications

Whenever client companies receives resumes with their requirements it should have some ranking. Database of system having bulk of resumes. But companies also sometimes tried to connect with candidate's with their social profiles like Github, Linkedin for more information [20].

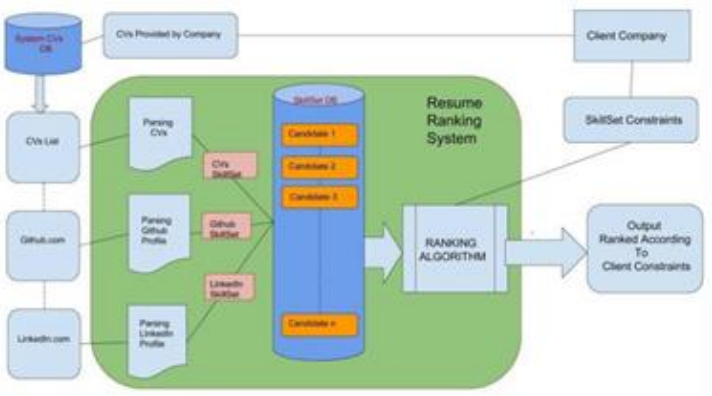

Fig.9. General architecture of resume parsing system

Bichitra Mandal et.al[7] discussed use of HDFS(Hadoop Distributed File System) with Map Reduced model for counting of number of consecutive words in Word processor. Hadoop is based on a programming paradigm called Map Reduce. It consists of a single master job tracker and one [6] slave task tracker per group node. The master is accountable for setting up the jobs to the slaves, screen them and re-execute them when the assignment fails. The slaves implement the everyday jobs as aimed by the master. The Map Reduce framework 10 works on key-value pairs and generates a set of key-value pairs as output.

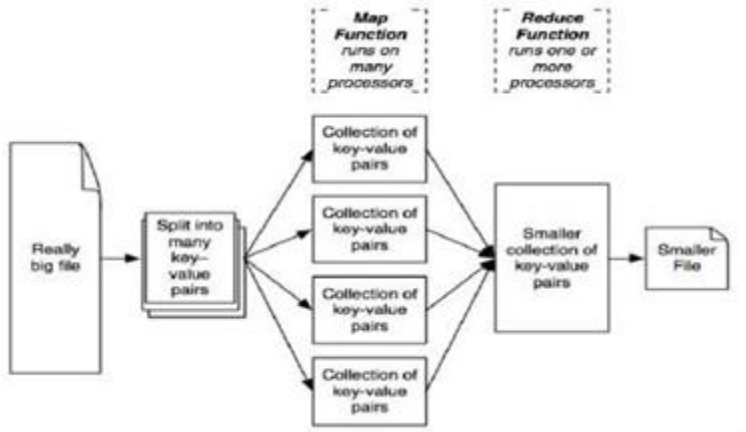

Fig.10. MapReduce Framework

3) Architecture for Word Processing using Hadoop Map Reduce:

A MapReduce job separates the key data set into self sufficient chunks which are later on refined by the map tasks coordinately. The input and output of the work are [6] stored mutually in a file-system. It handles the scheduling tasks, monitors them and enforces the unsuccessful tasks. For example: word processing using Word Count. In fig the file system is Map Reduced for efficient word processing [6] using Word Count. It reads text files and counts the number of occurrences of each word in a given input set. The mapper function takes each line as input and divides it into words, producing a key-value pair of each word. The output produced by the [6]mapper is then shuffled and sorted by value as per the occurrences of each word, which then becomes input to the reducer. Then the reducer sums the counts of every word and generates a single key-value with the word. The reducer also plays the role of a combiner on the map outputs. It minimizes the quantity of data sent by associating each word into a distinct record. HDFS stores both the input and output of the file system. If the input is not present in HDFS and is in local file system then the data is first copied into HDFS. The mapper processes one line at a time. It separates the lines into tokens and then produces a key-value pair of each word. The output of each map is then shuffled and sorted and approved by the combiner, which is also known as Reducer. The Reducer sum up the values for each key and the total output is generated as shown in fig 11.

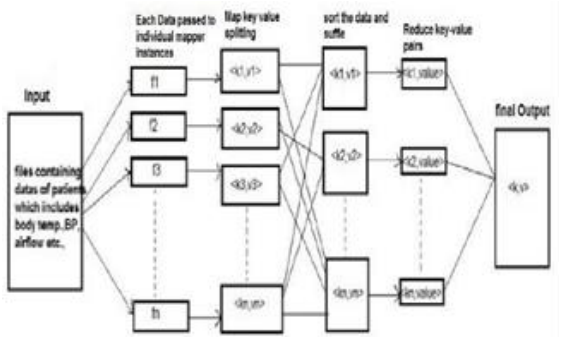

Fig.11. Word processing using hadoop MapReduced with one input System

Authors Pooja et al. [21] discussed the resume parsing application for transforming the resumes uploaded in different formats i.e doc, docx, pdf, text etc. The system transforms the uploaded resume into the desired format with only necessary details of candidates. The resume parser application is 24x7 available and easily accessible. This application finds out el-igible candidates for the given job position on the basis of uploaded resume. The authors describe the differences between the existing system and their proposed system. In existing system candidate has to manually fill the job profile. Although this process is lengthy and time consuming. Digitized resume format is not given, job recruiter has to check every parameters for searching the eligible candidate. So, a system is developed shown in fig 12,where candidate only upload their resume and details will automatically fetched [22].

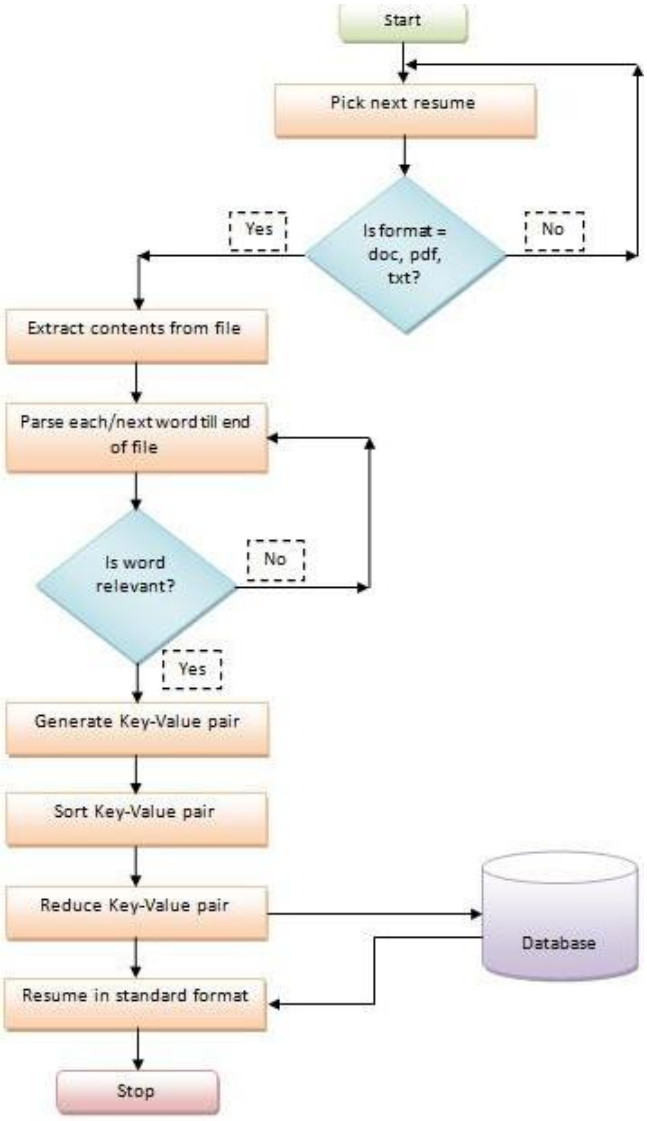

Fig.12. Flowchart of proposed resume system

The system shown in fig 12 having three modules. candidate module, job recruiter module, resume parser and standard-ized module. Candidate module helps candidate to find their appropriate jobs for applying and uploading resumes. In Job recruiter module, recruiter can view the standardized resume format passed by the system and can select eligible candidates. Resume parser and standardized module convert the original resume into a standard format with some classify-ing attributes like name, phone no, education, technical skills etc. Third module standardize the resumes on First In First Out(FIFO) basis. The job recruiter can view both the resume formats; the original one uploaded by candidate & standardized one. The advantage of this module is, it suggests the recruiters about eligible candidates. Recruiters don't need to search manually. For parsing the resume in the system it follows 3 approaches. Metadata, Natural Language Processing(NLP) & hybrid [21].

Metadata: This approach works in a fixed domain value format. Data will be stored in database table. The fixed values may be skills, gender, education etc. Parser will match with resume, and when match found key-value pair is generated.

Natural Language Processing(NLP): It works in fixed format such as email-id, mobile number, date etc. Each word passed by the parser will match up on basis of their meaning & if match is found, word will be tagged and then key-value pair is generated.

Hybrid Approach: It is the combination of other two approaches i.e metadata and NLP approach. This approach uses these two for extracting details.

The authors [22] researched on Chinese resume document analysis. It is based on pattern matching, feed-back control algorithms etc. This parsing is on semi-structured document. And the system was developed for china HR, which is biggest recruitment website. They used the resume parsing method for information retrieval on semi-structured Chinese document as shown in fig 13.

Name: Zhang Wu-ji
Gender: male **Birth date**: 1983-9 Mobile: 1381135****
E-mail: jiangzhixiang**@gmail.com **Address**: BUPT 29#, Haidian District
Education
2006-9 ~ now
University: Beijing University of posts and telecommunications
School: School of Telecommunication
Major: computer application **Degree**: master
Work Experience
ZTE Corporation. Time 2008-6 ~ Now
Job description Learn VLAN module, SVLAN module, LACP module, BFD module in Ethernet switches. Test the OAM module.
Tool C language, Source Insight. Rational purify, Rational purecoverage, CUnit.
Project Experience
Hebei Mobile Marketing System Time 2007-9 ~ 2008-2
Item description The system is designed for HeBei mobile company. You can handle kinds of businesses, the call automatically, SMS (by SMS platform) in the mobile terminal.
Responsibility I am in charge of developing Proxy Server, Application Server, Simulation

Fig.13. Chinese resume sample

The developed system has three main parts. Resume text class set, identifying algorithm, system project. Class set of resume text divides in simple items& complex items. Simple items include name, gender, birth, location, email etc and complex item includes learning experience, work experience, project experience, training experience, skills, and incentives as shown in fig14.

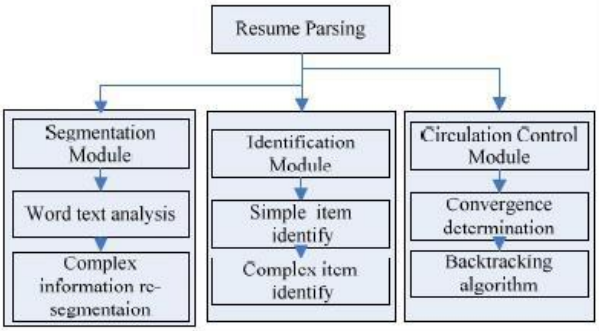

Fig.14. Methods arrangement of chinese resume parser system

For complex information, system uses vector space model based text automatically categorization method. They have used 5000 Chinese resumes as experimental data. Between those 2000 resumes used as training data & 3000 resumes used as a test samples. They uses accuracy as test indicator. If N1=Number of right information N=Total number of information

Then, Accuracy= N1/N.

They proved that, high accuracy is gained when information extraction was based on regular expressions and text automatic classification. And in complex dataset extraction, fuzzy match-ing algorithm is used for improvement in accuracy.

Table 2. Related implemented work

Paper Title	Techniques/Tools	Objectives	Challenges
Architecture of Efficient Word Processing using Hadoop Map Reduce for Big Data Applications [6]	Hadoop Map Reduce, Hadoop Distributed File System	To count the number of consecutive words and repeating lines	Time Consuming Method
Resume Parsing and Standardiza-tion [21]	Natural language processing	To find an eligi-ble candidate for a job on the basis of uploaded re-sume	It is limited on certain format extrac-tions,other formats can also be added.
Resume Parser: Semi-structured Chinese document anlysis [22]	SVM classification model	Extract informations from semi-structured Chinese resume format	Limited comparable dataset.

V. Proposed CV Parser Model

This section discusses about the proposed CV parser model for extracting information. Whenever the Jobseekers start up-loading their Resume/CV in recruitment website, CV parser extract data items for uploading CV to extract skill code like Academic background, Personal information etc.CV uploading can be done digitally or manually. Whenever the uploading is manual it is digitized by the process of digitization. Then the parser analyzer analyze all the information and extract it in the form of parsing information. The information collected after parsing is used for further decision making of the recruiters. Below is the model for the CV parser. The main job starts after the digitization process.CV

parsing tool extract all the relevant data from the parser box which includes students academic information, personal information. Job recruiter companies use these method for selecting the best CV. Uploading CV in a job portal is very easy. It reduces the hurdles and thus simplify the procedures. Incoming application extracts name, title and relevant features by parser for application management systems.

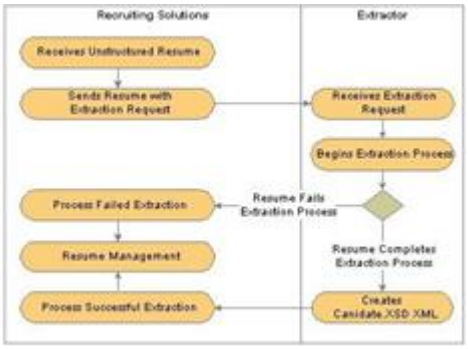

Fig.15. CV parsing Model

Above model in fig 15 having different components. The whole process consists of two parts. One is Recruiting Administrator and other is Recruiting solutions. Recruiting administrator gathers resumes in electronic or digitized manner for loading the resumes. Afterwards working of recruiting solutions starts. As the resume loading starts, extraction process also starts. This extraction process is controlled by Resume Management process. Resume management manages both successful extraction and unsuccessful extraction. It loads the resumes in the both cases. After successful loading of resumes, resume screening method is applied. Resume screening eliminates the candidates whose CV does not meet the job requirements.

A. *Algorithm for CV Parser Model*

Algorithm 1 Algorithm for CV Parser

Model Input: Client Request

1: J_T ! f(N)

2: f_t = f8T_1; T_2; T_n J_tg Allocated to ! fD_1; D_2; D_ng

3: f_c= Count of Tasks & f_p=Process Of Tasks

4: Integrating tasks from f_t

Output: Job J_T is complete.

Client request is the request sent to namenode.J_T is the total job and f(N) is the function in which J_T is described & it also stands for the function of the assignment of job to the Name node. f_T is the function where the tasks T_1, T_2, T_n belongs to the total job J_T .Function f_c is the number of tasks counted. And f_p is the process of tasks. Function f_c and f_p does the job of uploading CV either it is digitally or manually. Afterwards tasks are integrated from job f_T .Then the job J_T is complete.

VI. Implementation of Resume Parser Model

Resume parsing is extraction of CV or Resume in a free form document which is a structured information and the output generated will be in XML format. It will be suitable for storage and can be easily manipulated by user. Recruitment agencies work with CV/Resume Parsing tools to automate the storage and analysis of CV/Resume data. This saves recruiters hours of work by eliminating manual processing of each job application and CV they receive. The most common CV/Resume format is MS Word. Despite being easy for humans to read and understand, is quite difficult for a computer to interpret. Unlike our brains which gain or disseminate context through understanding the situation along with taking into consideration the words around it, to a computer a resume is just a long sequence of letters, numbers and punctuation. A CV parser is a program that can analyze a document, and extract from it the elements of what the writer actually meant to say. In the case of a CV the information is all about skills, work experience, education, contact details and achievements. Recruiters use resume parsing to create a far more convenient and efficient resume and application screening process. Resume parsing helps recruiters out in a huge way. This technology allows recruiters to electronically gather, store and organize the information contained in resumes or applications.

A. *Extraction of Entity*

We can use NLP(Natural Language Processing) for parse out a text into paragraphs and sentences. This kind of text analysis is difficult in other programming languages. Because human language can be rich So that computer languages can not capture & encode the amount of information. So, R and python are best for analysis [23] of data. Because R is having good libraries for Natural language Processing.

R can interface with other languages like C, C++, Java. Writing code in R having advantage of its functional programming style and its many other libraries for data analysis. Indeed most of the techniques such as word and sentence tokenization, n-gram creation, and named entity recognition are easily performed in R. Below is the extraction of names of people and places from a sample document.

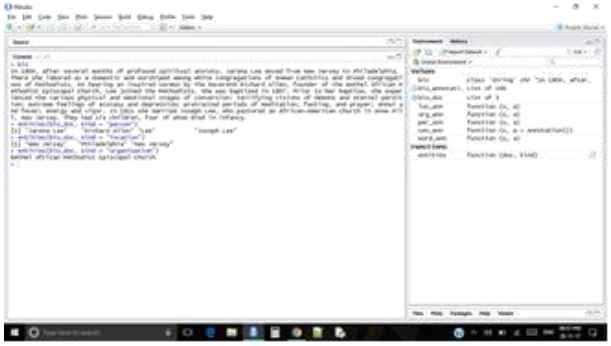

Fig.16. Extraction of person,location and organization from the text file in R

B. Annotation process

We need to create annotators for words and sentences. Annotators are created by functions which load the underlying Java libraries. These functions then mark the places in the string where words and sentences start and end. The annotation functions are themselves created by functions.

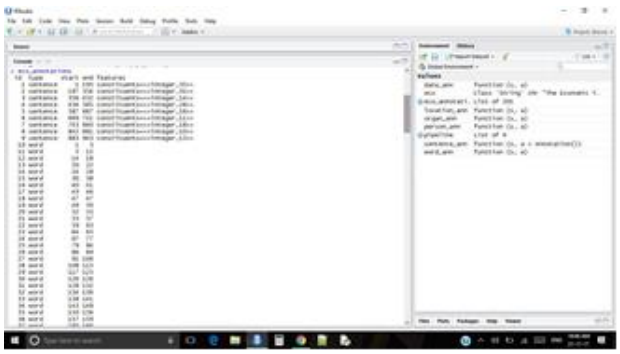

Fig.17. Sentences and word annotations

C. Tokenization

This is the very first step of text pre-processing [24] method. We have taken a text file in R and breaks into words and sentences. These are called tokenization, as we are breaking up the text into units of meaning, called tokens.

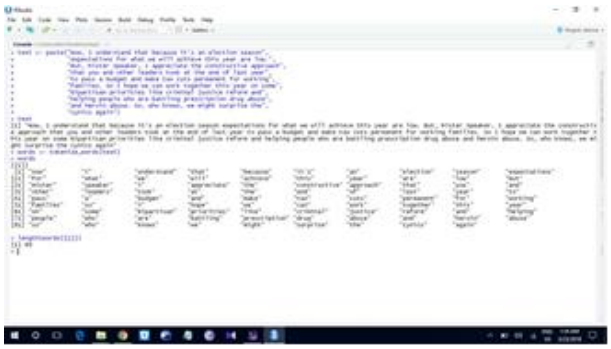

Fig.18. Tokenization of each word of sentences and its length

Now, we are taking one text as an input and tokenized it as a character vector, one-dimensional R object consisting only of elements represented as characters. It is possible to pair the output of the sentence tokenizer with the word tokenizer. If we pass the sentences split from the paragraph to the tokenize words function, each sentence gets treated as its own document. Apply this using the following line of code and see whether the output looks as you would have expected it, using the second line to print the object. We can also calculate the length of every sentence in the paragraph with one line of code.

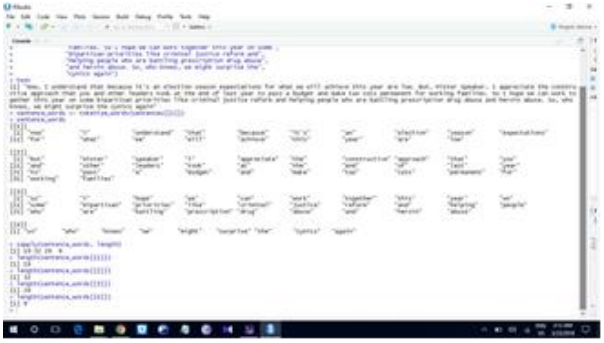

Fig.19. Detection of sentence boundaries

D. Counting frequent words

Frequent words or stop words can be counted. As example ”Hello” and ’hello’. Below is a graph for counting frequent word from a document.

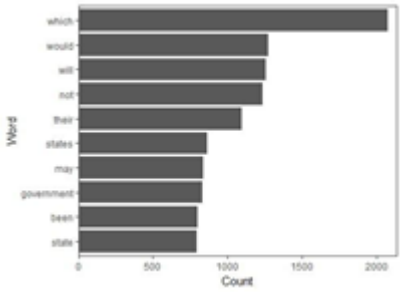

Fig.20. Frequent word count

E. Keyword searching through R

We are taking one resume as a input and search in base of different keywords.

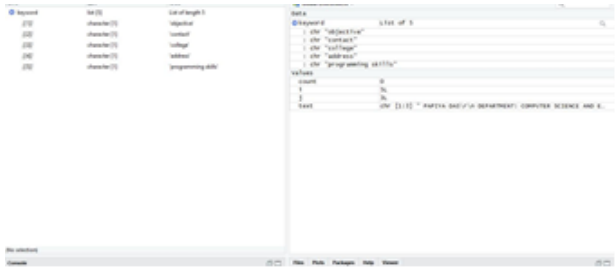

Fig.21. Keyword search

F. POS tagging

Part-of-speech tagging (POS tagging or PoS tagging or POST), also called grammatical tagging or word-category disambiguation, is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its contexti.e, its relationship with adjacent and related words in a phrase, sentence, or paragraph. A simplified form of this is commonly taught to school-age children, in the identification of words as nouns, verbs, adjectives, adverbs, etc.

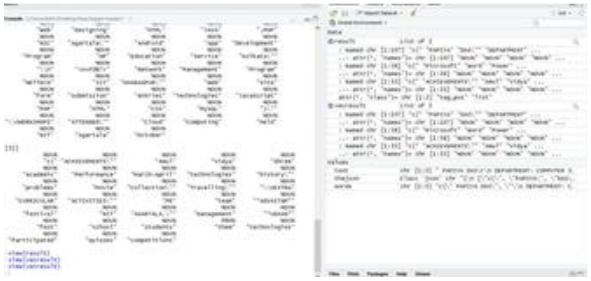

Fig.22. POS tagging through R

VII. Conclusion

The data sets of big data are large and complex in nature. In big data, text analysis has different aspects i.e entity extraction, concept extraction, article extraction, microformat extrac-tion etc. Now, everyone seeks jobs through online job por-tal by uploading resumes. All the resumes are not in struc-tured format. Some resumes are unstructured and also semi-structured. Our problem definition is based on designing an automated resume parser system, which will parse the uploaded resume according to the job profile. And it will transform the unstructured resumes into structured format. It will also main-tains a ranking system on the resumes. Ranking will depend on the basis of information extracted i.e according to technical skills, education etc. So, many software have been introduced to tackle such large databases.CV parsing is such a technique for collecting CV's.CV parser supports multiple languages, Semantic mapping for skills, job boards, recruiter, ease of customization. Parsing with hire ability provides us accurate results. It's technology increases the speed for ordering resume according to its types and formats. Its integration makes users API key for integration efforts. The parser operates using some rules which instructs the name and address. Recruiter com-panies use CV parser technique for selection of resumes. As resumes are in different formats and it has different types of data like structured and unstructured data, meta data etc. The proposed CV parser technique provides the entity extraction method from the uploaded CV's. The future scope of work is to implement and provides a brief analysis in real time database to perform the analysis with the existing models.

References

[1] H. Joshi and G. Bamnote, "Distributed database: A survey," Interna-tional Journal Of Computer Science And Applications, vol. 6, no. 2, 2013.

[2] R. Narasimhan and T. Bhuvaneshwari, "Big dataa brief study," Int. J. Sci. Eng. Res, vol. 5, no. 9, pp. 350–353, 2014.

[3] A. Halavais and D. Lackaff, "An analysis of topical coverage of wikipedia," Journal of Computer-Mediated Communication, vol. 13, no. 2, pp. 429–440, 2008.

[4] J. A. Stankovic, "Misconceptions about real-time computing: A serious problem for next-generation systems," Computer, vol. 21, no. 10, pp. 10–19, 1988.

[5] M. Ferguson, "Architecting a big data platform for analytics," A Whitepaper prepared for IBM, vol. 30, 2012.

[6] B. Mandal, S. Sethi, and R. K. Sahoo, "Architecture of efficient word processing using hadoop mapreduce for big data applications," in Man and Machine Interfacing (MAMI), 2015 International Conference on. IEEE, 2015, pp. 1–6.

[7] S. Vijayarani and M. R. Janani, "Text mining: open source tokenization tools–an analysis," Advanced Computational Intelligence, vol. 3, no. 1, pp. 37–47, 2016.

[8] R. Gaikwad Varsha, R. Patil Harshada, and V. B. Lahane, "Survey paper on pattern discovery text mining for document classification."

[9] "Entity extraction— aylien," http://aylien.com/text-api/ entity-extraction.

[10] "Named-entity recognition - wikipedia," https://en.wikipedia.org/wiki/ Named-entity recognition.

[11] "Entity extraction: How does it work? - expert system," www. expertsystem.com/entity-extraction-work/.

[12] "Named entity extraction — lexalytics," https://www.lexalytics.com/ technology/entity-extraction.

[13] Botan, R. Derakhshan, N. Dindar, L. Haas, R. J. Miller, and N. Tatbul, "Secret: a model for analysis of the execution semantics of stream processing systems," Proceedings of the VLDB Endowment, vol. 3, no. 1-2, pp. 232–243, 2010.

[14] T. Garcia and T. Wang, "Analysis of big data technologies and method-query large web public rdf datasets on amazon cloud using hadoop and open source parsers," in Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on. IEEE, 2013, pp. 244–251.

[15] Z. Jianqiang and G. Xiaolin, "Comparison research on text pre-processing methods on twitter sentiment analysis," IEEE Access, vol. 5, pp. 2870–2879, 2017.

[16] D. C ̧elik, A. Karakas, G. Bal, C. Gultunca, ¨A. Elc ̧i, B. Buluz, and M. C. Alevli, "Towards an information extraction system based on ontology to match resumes and jobs," in Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual. IEEE, 2013, pp. 333–338.

[17] M. Jose, P. S. Kurian, and V. Biju, "Progression analysis of students in a higher education institution using big data open source predictive modeling tool," in Big Data and Smart City (ICBDSC), 2016 3rd MEC International Conference on. IEEE, 2016, pp. 1–5.

[18] F. Javed, Q. Luo, M. McNair, F. Jacob, M. Zhao, and T. S. Kang, "Carotene: A job title classification system for the online recruitment domain," in Big Data Computing Service and Applications (BigDataSer-vice), 2015 IEEE First International Conference on. IEEE, 2015, pp. 286–293.

[19] W. Hua, Z. Wang, H. Wang, K. Zheng, and X. Zhou, "Understand short texts by harvesting and analyzing semantic knowledge," IEEE transactions on Knowledge and data Engineering, vol. 29, no. 3, pp. 499–512, 2017.

[20] "Intelligent hiring with resume parser and ranking using natural ..." https: //www.ijircce.com/upload/2016/april/218 Intelligent.pdf.

[21] P. Shivratri, P. Kshirsagar, R. Mishra, R. Damania, and N. Prabhu, "Resume parsing and standardization," 2015.

[22] Z. Chuang, W. Ming, L. C. Guang, X. Bo, and L. Zhi-qing, "Resume parser: Semi-structured chinese document analysis," IEEE, pp. 12–16, 2009.

[23] Ulusoy, "Research issues in real-time database systems: survey paper," Information Sciences, vol. 87, no. 1-3, pp. 123–151, 1995.

[24] A. M. Jadhav and D. P. Gadekar, "A survey on text mining and its techniques," International Journal of Science and Research (IJSR), vol. 3, no. 11, 2014.

Authors' Profiles

Papiya Das is currently pursuing M.Tech (Computer Science and Engineering) at the School of Computer Engineering, KIIT University, Bhubaneswar. Her ar-eas of interest Data Analytics, Data mining etc. She can be reached at papiyanita895@gmail.com.

Manjusha Pandey Ph.D (Computer Science), Mem-ber of IEEE is Professor at the School of Computer Engineering, KIIT University, Bhubaneswar. She has more than a decade of teaching and research experience. Dr. Pandey has published numbers of Research Papers in peer-reviewed International Journals and conferences. Her areas of interest is WSN, Data Analytics etc. She can be reached at man-jushafcs@kiit.ac.in.

Siddharth Swarup Rautaray Ph.D (Computer Science), Member of IEEE is Professor at the School of Computer Engineering, KIIT University, Bhubaneswar. He has more than a decade of teaching and research experience. Dr. Rautray has published numbers of Research Papers in peer-reviewed International Journals and conferences. His areas of interest is Image Processing/DA/Human Computer Interaction. He can be reached at siddharthfcs@kiit.ac.in.

How to cite this paper: Papiya Das, Manjusha Pandey, Siddharth Swarup Rautaray, "A CV Parser Model using Entity Extraction Process and Big Data Tools", International Journal of Information Technology and Computer Science(IJITCS), Vol.10, No.9, pp.21-31, 2018. DOI: 10.5815/ijitcs.2018.09.03