ONLINE RESUME PARSING SYSTEM USING TEXT ANALYTICS

Divyanshu Chandola[1], Aditya Garg[2], Ankit Maurya[3], Amit Kushwaha[4]

[1] *Student, Department of Information Technology, ABES Engineering College, Uttar Pradesh, India*
[2] *Student, Department of Information Technology, ABES Engineering College, Uttar Pradesh, India*
[3] *Student, Department of Information Technology, ABES Engineering College, Uttar Pradesh, India*
[4] *Student, Department of Information Technology, ABES Engineering College, Uttar Pradesh, India*

Mentor – Mr. Ashwin Perti
Associate Professor, Department of Information Technology, ABES Engineering College, Uttar Pradesh, India

Abstract

The task of finding the right candidate for a particular job can be a very tiring task for the HR department of an organization. Going through hundreds of resumes is not an easy task. No one has enough time to go into the details of any resume. This may result in shortlisting of wrong candidate or rejection of a right candidate which may result in significant loss of money and other resources. To simplify this process, we propose a Text Analytic approach to judge resumes on the basis of their content. Sentiment Analysis approach can also prove vital to analyze a candidate's resume on the basis of the description he or she provides. Sentiment Analysis is being used in various scenarios like recording people's responses for services and products. But here, we will follow that approach to judge a job candidate's resume. With the help of this method, we can help the employer to identify which candidate suits best the requirements of the company.

Key Words: *Text Analytics, Sentiment Analysis, Natural Language Processing, Text Mining, Lexical Analysis*

----------***----------

1. INTRODUCTION

This document is Text Analytics can be defined as a set of statistical, linguistic and machine learning techniques which let us analyses textual content in a structured manner so that it can be used for deriving higher quality information from unstructured data. It is also referred as Text Mining [1].The process involves structuring of text, using it to derive different patterns and evaluating them to get some useful output from it. Various methods of Natural Language Processing (NLP) are involved in Text Analysis like Lexical Analysis, Pattern Recognition, Information Retrieval, Data Mining, Parsing, Sentiment Analysis and Information Extraction. All these techniques help in enabling the computers to understand human language and analyses it like a human.

Sentiment Analysis or Opinion Mining [2] aims to determine the attitude of a speaker or a writer with regard to anything he or she has said or written. It tells what a particular person is trying to communicate, his or her emotional state and judgment regarding any topic. In this process a given text is taken as input and the words and sentences found in the document are categorized into different levels of sentiments. For example, words like 'Happy', 'Cheerful' describe Positive emotion and words like 'Sorrow', 'Sad' describe a negative emotion. Basic approaches in Sentiment Analysis involve keyword spotting, lexical affinity, latent semantic analysis, support vector machines and concept level approaches which use ontologies and semantic networks.

Resumes are a great source of unstructured data which can be usefully analyzed by the companies to shortlist the right candidate. Various qualities of a candidate can be identified based on the content of his resume. Just like humans, a computer can analyses the resume by finding the right keywords which will categorize the level of every candidate on a scale of 3, Low, Average and High. Initially, a training data set would be manually created so that the computer is able to identify what characteristics make a candidate's resume better or worse than others. A learning algorithm can be created which would extract useful keywords from each and every resume which will be analyzed by the system.

Learning technique used can be either Supervised or Unsupervised. Supervised learning would involve training data set for each class of levels defined on the scale. The techniques involved in classification under supervised learning are Support Vector Machine, K-nearestneighborr and Naïve Bayes. Unsupervised learning don't use any training set data rather the use of clustering algorithms like K-means clustering can be used to classify data into various categories or levels. Semantic Orientation is also a very efficient technique for classification.

The paper is divided in following sections – Section 2 describes the related work which has been done. Section 3 will present the detailed explanation of the methodology we propose to adopt. Section 4 shall give an insight of the prototype we have implemented till now. Section 5 shall present the conclusion and the proposed future work which can be done.

2. RELATED WORK

Resume RDF ontology has been introduced by Uldis Bojars and John G. Breslin [3] which uses an RDF data model to model a resume. ResumeRDF describes resume information with its lavish set of classes and properties. Uldis Bojars further extended FOAF with resume information [4] for an even more improved description of information.

In 2002 and 2003, Turney and Littman proposed a strategy which would infer the semantic orientation or evaluative character of a word from its huge hundred billion-word corpus corpora taking into consideration the semantic associations with the other words, referred as paradigms by him. [5][6].

Ujjal Marjit et al. [7] proposed a different technique which retrieved resume information using the concept of Linked Data enabling the web to share data with different sources enabling it to discover multiple kinds of information. An ontology based approach was proposed by Maryam Fazel-Zarandi et al. [8] which would match job seekers skills with the help of a deductive model which determined a match between the skills of a job seeker and the skills required by the recruiter.

Another system to automate resume information extraction was developed by Kopparapu of TCS Innovations lab [9] which featured rapid search of resume extracting useful information from a free format resume with the help of various NLP techniques.

Online Chine resume parser was presented by Zhi Xiang Jing et al. [10] which used rule based and statistical algorithms to extract information from a resume. Zhang Chuang et al. [11] worked on a resume document block analysis which was based on pattern matching and multi-level information identification making the biggest resume parser system.

Celik et al. [12] designed a system which converted a resume into an ontological structural model which simplified the analysis of Turkish and English resumes. Di Wu et al. [13] managed to extract information from resumes more effectively by the concept of ontology using WordNet for similarity calculation.

Although there are many other existing websites which provide advanced facilities like searching on the basis of keywords, domain, location etc., their search does not take into consideration, the skill level of a particular candidate. For example, if a company searches for a candidate who can work in C language, they can easily search for candidates who have C language mentioned in their resumes. But how will they know the proficiency of that particular candidate in C language.

In our prototype, we use extra information like the projects in which the candidate was involved as well as the project description. This information will be taken as input from the candidate and by analyzing the text used by him or her; we will categorize the candidate into various expertise levels. So if a company wants an employee who necessarily has a high expertise level in C language, only then his resume will be shortlisted. A knowledge base of various keywords will be designed which will form the basis of categorization. This will also help in ranking of various resumes to tell which one is better or worse than the other giving the applying candidates a chance to present themselves in the best possible way.

3. METHODOLOGY

The model we propose has four steps:

- Collection of resumes.
- Searching for keywords stored in knowledge base in the resume text.
- Fetching new keywords from the resumes to build the knowledge base further.
- Ranking and Categorization of candidate based on a rating score.

Figure (1) shows the proposed model and all the steps are explained in subsequent sub sections.

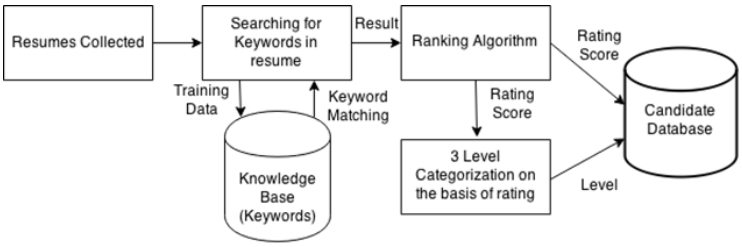

Figure 1: Proposed Model

3.1 RESUME COLLECTION

This step involves the collection of various resumes uploaded by the candidates. A simple web interface has been designed in our prototype model which will make the user fill a form having the fields which would be required to be filled by the job seeker. Our prototype deals with candidates for IT companies but this can be generalized for various other sectors by using an even more extensive knowledge base.

The candidates will specify the languages they know along with the projects on which they have worked. This will help the hiring company as they can easily filter out the candidates who do not have the knowledge of the language which is demanded by the company. Most websites use this as their filter method by searching with a keyword. For example, if they want a candidate who knows Java, they can simply search for 'Java' in the resume to filter out candidates who do not know Java. But this technique does not tell the company anything about the proficiency level of the particular candidate in the language he or she knows. There is no way to tell how good the candidate is in Java.

3.2 KEYWORD SEARCHING

This is one of the most crucial steps of our model. A knowledge base consisting of various keywords is made from the initial training data. The input text which is received needs some pre-processing before it can be used. For this purpose, we use a POS tagger and a chunker which are used to split the text into sentences, which are then analyzed by a syntactic parser which labels all the words with their part of speech information. Using a chunker helps in providing a flat structure of extracted data [14]. Lexical Analysis [15] can also be done to tokenize the words which can be then categorized for the purpose of parsing.

The keywords are extracted from the analyzed set of words. The nouns, verbs and adverbs are the part of speech tags which are targeted for extraction while others can be dropped.
Extracted words are then compared with the keywords stored in the knowledge base. Every word stored in the knowledge base has a value associated with it. These values are defined based on the importance of the word. Since our prototype deals only with resumes for jobs in IT companies, we have used various keywords which are extracted from the description of projects in which the candidate was involved. A large set of valued keywords can be made and used to rate the candidates on the basis of words extracted from their resumes. The sum of all the keyword values is calculated to obtain a rating score which will be used further to rank the resume and categorize the candidate on the basis of rating.

3.3 ADDITION IN KNOWLEDGE BASE

While the keywords found in resume text will be matched, the words which are not found in knowledge base are further analyzed and if found relevant, is added to the knowledge base. Since the data from which knowledge extraction has to be done is unstructured, we follow traditional methods of information extraction. Apart from that, Ontology based Information extraction can also be done by Semantic Annotation [16] in which we augment the natural language text into metadata which can be represented in form of RDFa (Resource Description Framework in attributes) [17]. The process is divided into two subtasks – Terminology Extraction and Entity Linking.

For terminology extraction, domain specific lexicon can be used after tokenizing the text. After that, a link is created between extracted lexical terms and the concepts from either ontology or the already existing knowledge base. Lastly, the context of the various terms is analyzed so that they can be correctly assigned to the level in which they should belong. In this way, knowledge base can be regularly updated and also, it will be manually examined regularly to remove keywords which may no longer be useful.

3.4 RANKING AND CATEGORIZATION

After getting the rating score of the resume, a candidate can be ranked on the basis of his resume's score. This will be useful in comparing two candidates while shortlisting them. Whenever the company searches for a candidate keeping in mind certain requirements, the candidate who is ranked above will be presented to the company first which would be adding to his advantage in cases where the vacancies available may not be high.

More important procedure which has to be followed is of categorization. The sentiment analysis categorizes the people's opinions as Positive, Negative or Neutral to derive results. Similar to that, our model would categorize candidates as Low, Average or High on the basis of their resume.

In our prototype, we have categorized the resumes of candidates applying for IT companies in the same 3 level scale and considered it as their expertise level in the programming language mentioned in their project description which would help the company shortlist only those candidates whose expertise level in a particular language is as required by the company.

In this way, the efficiency of recruitment process of a company could be significantly improved as better candidates would be picked up without needing to give a lot of time in going through the resumes manually.

4. IMPLEMENTATION

The algorithms for matching of keywords have been implemented on Python facilitated by MySQL connector which fetches data required for matching from the table of Keywords and their associated values which form the Knowledge Base. The algorithm matches the extracted keywords with the keywords present in the knowledge base and stores them in a different list along with their rating values. The rating scores of individual keywords after being added are returned to the candidates table for the purpose of their ranking on the basis of score. Categorization is performed on the basis of rating score of each candidate. The rules for rating and categorization followed in the prototype are as follows -

Rating scale for individual keywords – 1: Low 2: Average 3: High

Rating Score = Sum of ratings of all keywords matched

Categorization on the basis of Rating Score – Below 10: Low 10 to 20: Average Above 20: High

The candidates and the company will use a website based interface to interact. Both of them after getting registered as users shall be added in the database, separate for candidates and companies. The candidate database consists of various fields the candidate would have to fill in while registering which includes the programming languages known and projects in which the candidate has been involved along with its description. These are the crucial fields which will

be used to determine the expertise level of the candidate. The fields for expertise level and rating score shall be automatically filled for every candidate once the resume is analyzed.

On the other side, a company can simply specify their requirements while searching for a candidate as shown in figure 2. They may specify the language or languages which should be known and the expertise level of a candidate which they need. Once they submit their requirements, the candidates who fulfil the criteria are fetched from the Candidates table and displayed to the company as shown in figure 3.

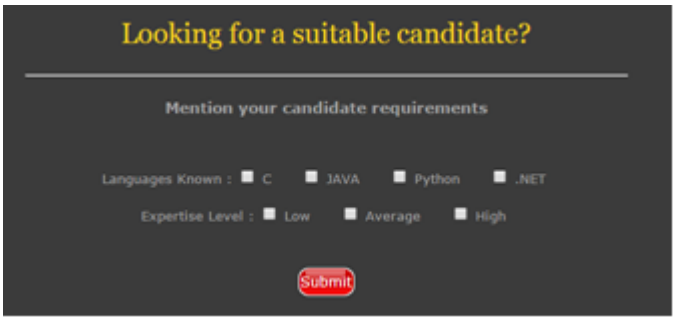

Figure 2: Candidate Requirements

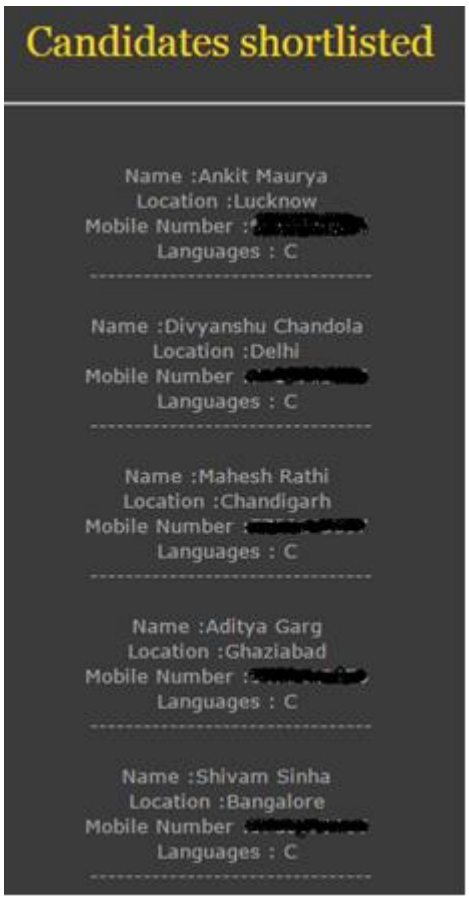

Figure 3: Candidates shortlisted

5. CONCLUSIONS

The model we propose efficiently shortlist the candidates according to the requirements of the company based on their resumes. Although one can question the trustworthiness of a resume to shortlist a candidate but since this will not be the final procedure of any company's recruitment process, it still holds its importance. Resumes are always considered as the first impression of any job seeker so it is important that the candidates focus on the way they describe themselves in resumes to get shortlisted by the company for further process. Candidates will get a chance to get them ranked above others on the basis of projects he has been involved in as well as the way he describes it.

In future, we will try to generalize the concept which is till now limited to only IT sector. For that, different criteria would have to be formulated which would become the basis for categorization and ranking of candidates.

REFERENCES

[1] http://en.wikipedia.org/wiki/Text_Mining
[2] http://en.wikipedia.org/wiki/Sentiment_analysis
[3] Uldis Bojars, John G. Breslin, "ResumeRDF: Expressing Skill Information on the Semantic Web".
[4] Uldis Bojars, "Extending FOAF with Resume Information".
[5] Turney, P.D., Littman, M.: Unsupervised learning of semantic orientation from a hundred billion-word corpus. Technical Report ERC-1094 (NRC 44929), National Research Council of Canada (2002)
[6] Turney, P.D., Littman, M.L.: Measuring praise and criticism: Inference of semantic orientation from association, ACM Transactions on Information Systems (TOIS), vol. 21, no. 4, pp. 315--346 (2003)
[7] Ujjal Marjit, Kumar Sharma and Utpal Biswas, "Discovering Resume Information Using Linked Data", in International Journal of Web & Semantic Technology, Vol.3, No.2, April 2012.
[8] Maryam Fazel-Zarandi1, Mark S. Fox2, "Semantic Matchmaking for Job Recruitment: An Ontology-Based Hybrid Approach", International Journal of Computer Applications (IJCA), 2013
[9] Kopparapu S.K, "Automatic Extraction of Usable Information from Unstructured Resumes to aid search", IEEE International Conference on Progress in Informatics and Computing (PIC), Dec 2010.
[10] Zhi Xiang Jiang, Chuang Zhang, Bo Xiao, Zhiqing Lin, "Research and Implementation of Intelligent Chinese Resume Parsing", WRI International Conference on Communications and Mobile Computing, Jan 2009.

[11] Zhang Chuang, Wu Ming, Li Chun Guang, Xiao Bo, "Resume Parser: Semi-structured Chinese Document Analysis", WRI World Congress on Computer Science and Information Engineering, April 2009.
[12] Celik Duygu, Karakas Askyn, Bal Gulsen, Gultunca Cem, "Towards an Information Extraction System Based on Ontology to Match Resumes and Jobs", IEEE 37th Annual Workshops on Computer Software and Applications Conference Workshops, July 2013.
[13] Di Wu, Lanlan Wu, Tieli Sun, Yingjie Jiang, "Ontology based information extraction technology", International Conference on Internet Technology and Applications (iTAP), Aug 2011.
[14] "Learning and Knowledge-Based Sentiment Analysis in Movie Review Key Excerpts" Bj¨orn Schuller and Tobias Knaup.
[15] http://en.wikipedia.org/wiki/Lexical_analysis
[16] Jalaj S. Modha Prof & Head Gayatri S. Pandi Sandip J. Modha, "Automatic Sentiment Analysis for Unstructured Data"
[17] Ben Adida, Mark Birbeck "RDF in attributes"