Resume Information Extraction Framework

Anaswara R[1], Aswathy T[2]

[1]M. Tech. Student, Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India
[2]M. Tech. Student, Computer Science and Engineering, Sree Buddha College of Engineering, Kerala, India.

---***---

Abstract - *Digital communication reduces the time to send resumes. But the recruiter's work became complicated. Every company's one of the crucial function is hiring new individuals. For recruitment a pool of resumes a company gets for a job application are way higher than the number of person assigned to analyze it. In early days whenever a company looking to hire someone, they had to process resumes by hand. Resumes are semi-structured documents. Which means in each file contains varying information, such as different number of fields, different field names, different nested format or different field type. This makes it harder to parse. There is a need for Text Mining Model that filter keywords like experience, interest, qualification etc. Most approaches focus on parsing to get information from resume. Based on those extracted keywords various categories can be defined and resumes can be categorized, ultimately leading to better individuals being selected.*

Key Words: **Parsing, Text mining, Clustering, Classification, Semi-structured document.**

1. INTRODUCTION

Resumes, one kind of common semi-structured document [1], usually contain valuable structured data hiding in personalized expression. The data may have predefined specifications, but information each file contains might vary, such as the different numbers of fields, containing different field names, field types or different nested format. This makes parsing the document a little more complex than it might be. Classification for them or handling them requires the user to manually open the files, read its whole structure, select the interested information and close them. This manual labour scales linearly with the number of target fields. But when available, this information can be utilized as a relationship table that can be used to answer accurate queries or to perform data mining tasks.

In this project use resume information extraction [2] paradigm where the system makes a single data-driven pass over a handful of rule expression. It extracts resume fields just require putting the electronic document in the specified folder.

2. RELATED WORKS

In this section, review some previous works on group anomaly detection.

Extraction of the information [2] from resumes has been an important area of focus for a lot of researchers. Work on resumes generally includes information extraction parsers, classifiers, and natural language processors and data storage structures. There are many commercial products to resume data storage, information extraction and retrieval. Some of the commercial products include: Daxtra CVX, Sovren Resume/CV Parser, ALEX Resume parsing, Akken Staffing, and ResumeGrabber Suite. There is no complete product specification, algorithms and techniques available for resume information extraction.

Online Chine resume parser was presented by Zhi Xiang Jing et al. [3] which used rule based and statistical algorithms to extract information from a resume. In this presents a systematic solution of the information retrieval in online Chinese resume. Chinese resume contents have several expression and the structure of resume is complex. So this here applies rule-based and statistical algorithm to extract information.

Another author Zhang Chuang et al. [4] worked on a resume document block analysis which was based on pattern matching and multi-level information identification making the biggest resume parser system. Semi-structured Chinese document analysis is the most difficult task for complex structure and Chinese semantics. According to the generic characteristics of the semi-structured document and the specific characteristics of the resume document, the paper researched on resume document block analysis based on pattern matching, multi-level information identification and feedback control algorithms was also prompted. Based on the research, Resume Parser system was implemented for ChinaHR, which is the biggest recruitment website. It can read, analysis, retrieval and store the information automatically.

There are many other existing websites that provide advanced facilities like searching on the basis of keywords, domain, location etc., and their search does not take into consideration, the skill level of a particular candidate. 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 the recruiters know the proficiency of that particular candidate in C language.

3. SYSTEM DESIGN

The proposed system supports the resumes in pdf and docx format. In this paper, use DBSCAN algorithm [5] for clustering which cluster the information extracted from a resume. Various resume classifications are generated. For this Gradient Boosting Machine is used. DBSCAN algorithms are used to cluster textual data.

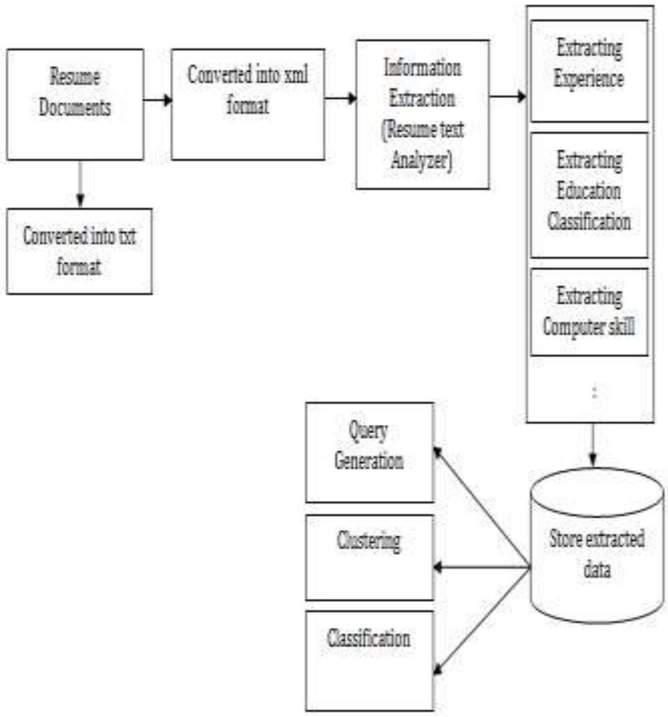
Resume Documents
Converted into xml format
Information Extraction (Resume text Analyzer)
Extracting Experience
Extracting Education Classification
Extracting Computer skill
:
Converted into txt format
Query Generation
Clustering
Classification
Store extracted data

Fig -1: Architecture of proposed system

1. Inputs are resumes, which are stored in a folder.
2. Resumes are converted to txt format.
3. Resumes converted into xml format.
4. Identify necessary data from the xml document and extract different keywords from resume.
5. The extracted data are stored into the table in the database. Rule expressions are used to extract data from resume.
6. Give score to education qualification, computer skills and language known.
7. Convert the text data into numerical value.
8. This numerical value is used for clustering. DBSCAN algorithm is used for clustering.
9. Then classify the resume for than GBM algorithm is used.

3.1 MODULES

The proposed system consists of one module:

- Admin Module

The main functionalities of the Admin module are:

Table -1: Admin module function and description

Function	Description
Upload Resume	Store the resume into a folder.
Convert Resume into .txt format	The uploaded resumes converted into .txt format and store it in a folder.
Convert Resume into XML format	The xml format of resume is created. This is used for extraction.
Extraction	Necessary fields are extracted from the xml format.
Query generation	Sql query is generated to search the resume.

4. IMPLEMENTATION

The proposed system has mainly five phases namely:

- ❖ Resume Conversion to txt format
- ❖ Taxonomy Creation
- ❖ Data Extraction
- ❖ Clustering
- ❖ Classification

4.1 RESUME CONVERSION TO TXT FORMAT

The uploaded resumes are either in pdf or docx format which are converted into txt format. Converting PDF files to plain text files—i.e. extracting text data from PDF-encapsulated files. With many Linux distributions, it is freely accessible and included by default, and is also accessible for Windows as part of the Xpdf Windows port. Use c# code to convert any DOCX files in Text files. Using this it will able to do what need*:*

- Create a fresh document and complete it with the required information.
- Load current document and get all of it structure as tree of objects.
- Modify the current document's paragraph formatting, text, tables, TOC and other components.
- Parse the document and pick up the tree of its objects.
- Replace, merge any data in documents and save data as new DOCX, Text or RTF.
- Convert between PDF, DOCX, RTF and Text.

4.2 TAXONOMY CREATION

Spire.Doc (Spire.Office) presents an easy way to convert Doc to Office OpenXML. In this way, convert an exist Word doc file to Office OpenXML format with a few clicks. When talking about Office OpenXML, may think of HTML. Office OpenXML is actually comparable to HTML, both languages are tag-based. The difference between Office OpenXML and HTML is that the tags which Office OpenXML uses are not predefined. If want to create own tags within Office OpenXML, need to follow a few rules.

Firstly, only one root element is contained in Office OpenXML document. The root element is often used as a document element and appears after the segment of the prolog. Besides, all the Office OpenXML elements should contain end tags. Both start and end tag should be identical. Also, the elements can't overlap. What's more, all attribute values must use quotation marks and can't use some special characters within the text. After following the rules, the Office OpenXML document will be well formatted. To covert pdf to xml format SautinSoft code is used.

4.3 DATA EXTRACTION

For different types of documents use different parsing methods. These documents can be divided into two categories, documents in PDF format or Word format, according to their suffix of filename. The text of the documents grabbed by programming according to the characteristics of their, text content can be divided into simple and plain text and text in XML format.

The method of extracting this type of document is different from the simple text flow, and the mainly task is analyzing the tree structure level, and then combine the matching based on regular expression.

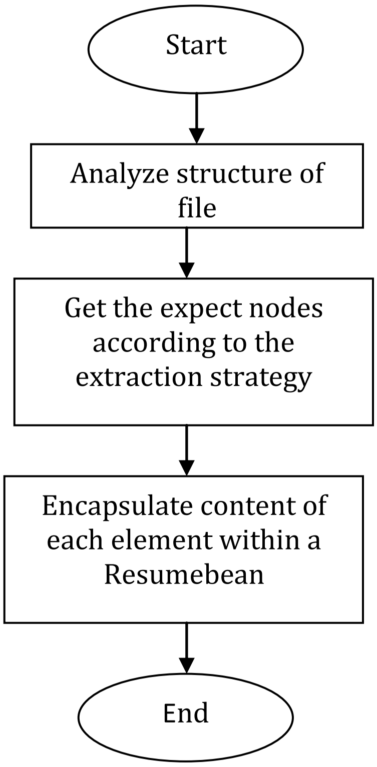
Start
Analyze structure of file
Get the expect nodes according to the extraction strategy
Encapsulate content of each element within a Resumebean
End

Fig -2: XML document parsing process.

To extract data from xml document of resume, first find all the <text> tag in the xml document. The <text> tag contains all the relevant data. This text tags are stored in a temporary notepad. Create a table which contains some necessary heads in the resume such as Educational Qualification, Experience, Computer Skills, and Language Known etc. Identify the content in between two heads that in the table. The stored this identified data to a table.

4.4 CLUSTERING

Clustering can be defined as the process of creating clusters. Each cluster is a collection of like-minded objects. It usually deals with finding a similarity in an unstructured collection of unlabeled data. Here use DBSCAN for clustering. The clustering is applied with the help of WEKA.

4.5 CLASSIFICATION

An increasing number of organizations have adopted text classification schemes to efficiently handle the ever-increasing inflow of unstructured data. The aim of text classification systems is to increase discoverability of information and make all the knowledge discovered available or actionable to help strategic decision making. Gradient Boosting Machine (GBM) algorithm is used for classification. In gradient boosting, it trains many model sequentially.

5. CONCLUSIONS

There are problems in resume processing and the selection of appropriate resumes from a large amount of resume. Resumes selected by choosing them based on defined job requirements and then highlighting their unique features. The skill categories, specific skills, and unique skills of a resume are considered to determine the uniqueness of a resume. This helps recruiters speed read through a set of resumes and their specialties in order to decide on prospective candidate. The resumes are selected based on the keyword extracted.

Clustering and classification improve the efficiency of the system. DBSCAN algorithm is used for resume clustering. DBSCAN algorithm is an efficient algorithm for text mining. For clustering first convert the extracted data into numerical value. Then generate the clusters of resumes.GBM algorithm is used for classification of resumes. This will generate different group resumes with similar properties.

The clustering and classification are implemented with the help of WEKA.

REFERENCES

[1] Jiang ZhiXiang, Identification of the Semi-structured text [D]. Beijing University of Posts and Telecommunications, 2009.

[2] Gong Yiguang, Mei Ping. Research on a Combined Ontology-based Text Information Extraction Technology [A]. Proceedings of 2010 International Conference on Broadcast Technology and Multimedia Communication (Volume 4) [C]. International Communication Sciences Association, Hong Kong: 2010:4: 129-132.

[3] 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.

[4] 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.

[5] Density-based clustering algorithms – DBSCAN and SNN by Adriano Moreira, Maribel Y. Santos and Sofia Carneiro.

[6] YAN Wentan, QIAO Yupeng, IEEE, "Chinese resume information extraction based on semi-structured text", 1934-1768, July 26-28, 2017.

BIOGRAPHIES

Anaswara R, she is currently pursing Master's Degree in Computer Science and Engineering from Sree Buddha college of Engineering, Elavumthitta, India. Her research area of interest includes the field Data mining.

Aswathy T, she is currently pursing Master's Degree in Computer Science and Engineering in Sree Buddha College of Engineering, Elavumthitta, Kerala, India. Her research area of interest includes the field of data mining, internet security and technologies.