International Journal of Advance Research in Science and Engineering
Vol. No.4, Special Issue (01), November 2015
www.ijarse.com
IJARSE
ISSN 2319 - 8354

ENHANCEMENT PROCESS OF ANALYZING UNSTRUCTURED DOCUMENT BY DEVELOPING TEXT MINING ALGORITHM

V.Jayaraj[1] P.Rajadurai [2]

[1]Associate Professor, [2]Research Scholar, Bharathidasan University, Thiruchupalli, (India)

ABSTRACT

Prevailing information systems allow companies to apprehend huge amounts of data. Much of this data captured is structured and can be analyzed using old-fashioned database software. Increasingly, however, plethora of textual data is unstructured, and confronts simple attempts to make sense of it. Manual analysis of this unstructured textual data is impractical, and as a result, numerous text mining methods are being developed to automate the process of analyzing this unstructured data. The primary objective of this paper is to propose a text mining algorithm to mine the data in resumes and enhance the scope of the recruitment process in firms. A novel algorithm named IRCF text mining is proposed to extract useful information from a set of resumes and employs a new technique named Weighted Ranking method during extraction where the information gained is relevant and efficient for business needs. 300 numbers of resumes are collected from timesjob.com for this experimental study. The experimental results showcased that the relevancy ratio of the extracted resumes are high and the execution time taken is comparatively low when compared with KNN algorithm. The most important aspect regarding resume mining is relevancy and accuracy in extraction of data and the proposed algorithm performed extremely well as weighted ranking method filters the irrelevant resumes effectively since the relevancy ratio almost matched with that of the manual relevancy calculation.

Keywords: Business Intelligence, Recruitment Process, Resume Mining, Text Mining.

I. INTRODUCTION

Data mining also known as Knowledge Discovery is the process of extracting or mining essential information from the vast amount of data. It is most commonly used in extracting information from ordered pattern. In major business, data are complex in nature and exist in deferent formats and often organized in a poorly manner (i.e.,) in an unordered manner. From these sources, the data mining alone cannot be efficient to extract the useful information. In order to extract the essential text from these textual documents, a new powerful tool has been used called Text Mining.

Text Mining is an art of acquiring potentially useful knowledge from the textual document. Data Mining cannot derive its impact on extracting useful details from large unstructured materials based on natural language. But Text Mining will be the solution. The process of text mining is also termed as Information Extraction or Information Retrieval or Document Classification. The process of extracting information from huge volumes of data is necessary in order to obtain proper knowledge from the useful information by snubbing unwanted

information. This useful information enables the administrator to arrive to the decision correctly and facilitates to improve their business to a greater extent. Most of the business records are maintained in the form of documents and hence the documents are in unstructured format. In order to obtain knowledge from this unstructured document requires more manual process and time. Hence to elude this snag, text mining plays a pivotal role in extracting the essential much needed information by categorizing the text. Thus, Text Mining can be also termed as "Categorization of Texts".

The globalization has lead to cut throat competition in business and the decision making is very dominant factor for the success of a business. Thus Business Intelligence is the process of developing the strategies in order to gain competitive advantage for business. To accomplish this, countless techniques have been emerged and utilized. One such technique that helps business intelligence is data mining. But data mining has certain limitations as it can be efficient in mining the information from the structured internal data and predicting the trends based on these data or knowledge to make wise decision. But some of the business handles the textual information which is not in a structured manner such as project report, employee resume, and competitor's profile.

These documents are presented in unstructured form, since the details are described in the form of text using natural language and each of the documents followed the own developers style. Unlike the data in the database, the information in the textual document are scattered through the text. From such kind of documents, the information can be mined with proper care. The afore said details can be implemented by the proposed algorithm after developing a configuration file to identify the useful information from the document, and various patterns can be easily extracted and organized in order to provide the meaningful information. Through the extracted information, the administrator can able to make the decision to enhance the business.

This paper is organized as follows: In section 2, various research works has to be analyzed in order to enhance our work. In section 3, the problem statements are discussed, in section 4 our proposed methodology has been described in details with proper implementation of the proposed algorithm. Following this, the experimental results of the proposed method are presented. Finally, the paper is concluded by summarizing the work.

II. RELATED WORK

Atika et al, in paper [1], they described that the textual data in electronic documents today around the world have no doubt brought forward all the information one could need and as data banks build up worldwide, and access gets easier through technology, it has become easier to overlook vital facts and figures that could bring about groundbreaking discoveries. The research paper discusses in detail an implementation of Information Extraction and Categorization in the text mining application that they had implemented. To extract terms from the document they had used modified version of Porter's Algorithm for inflectional stemming. For calculating term frequencies for categorization, they have used a domain dictionary for 'Computer Science' domain.

Garcia et al, in paper [2], they stated that extracting insights from large text collections was an aspiration of any organization aiming to take advantage of their experience generally documented in textual documents. Textual documents, either digital or not, have been the most common form to register any organization transaction. Free text style was a very easy way to input data since it does not require users any special training. On the other hand, the text material easily collected becomes the major challenge for building automatic deciphering tools. In

the paper they presented ADDMiner, a text-mining model for extracting causality relationships from a large text collection of accident reports. This model was based on using domain ontology as well as a corpus-based computational linguistics to guide the mining process. Examples from offshore oil platform accident reports illustrate the potential benefits of the approach.

Vaishali et al, in paper [3], described text mining technique for automatically extracting association rules from collections of textual documents. The technique called, Extracting Association Rules from Text (EART). It depends on keyword features for discover association rules amongst keywords labeling the documents. EART system ignores the order in which the words occur, but instead focusing on the words and their statistical distributions in documents. The system based on Information Retrieval scheme (TF-IDF) for selecting most important keywords for association rules generation. It consists of three phases: Text Preprocessing phase (transformation, filtration, stemming and indexing of the documents), Association Rule Mining (ARM) phase (applying the designed algorithm for Generating Association Rules based on Weighting scheme GARW) and Visualization phase (visualization of results). Experiments applied on Online WebPages related to the cryptography. The extracted association rules contain important features.

Raymond et al, in paper [4], they stated that Text mining concerns looking for patterns in unstructured text. The related task of Information Extraction (IE) was about locating specific items in natural-language documents. The paper presented a framework for text mining, called DISCOTEX (Discovery from Text Extraction), using a learned information extraction system to transform text into more structured data which was then mined for interesting relationships. The initial version of DISCOTEX integrates an IE module acquired by an IE learning system, and a standard rule induction module. In addition, rules mined from a database extracted from a corpus of texts are used to predict additional information to extract from future documents, thereby improving the recall of the underlying extraction system. Encouraging results are presented on applying these techniques to a corpus of computer job announcement postings from an Internet newsgroup.

Li Gao et al, in paper [5], they worked out the strong points of text mining in extracting business intelligence from huge amount of textual information sources within business systems. They will apply text mining to each stage of Business Intelligence systems to prove that text mining was the powerful tool to expand the scope of BI. After reviewing basic definitions and some related technologies, they discussed the relationship and the benefits of these to text mining. Some examples and applications of text mining will also be given. The motivation behind was to develop new approach to effective and efficient textual information analysis. Thus they expanded the scope of Business Intelligence using the powerful tool, text mining.

Divya Nasa, in paper [6], described that In today's world, the amount of stored information has been enormously increasing day by day which was generally in the unstructured form and cannot be used for any processing to extract useful information, so several techniques such as summarization, classification, clustering, information extraction and visualization are available for the same which comes under the category of text mining. Text Mining can be defined as a technique which was used to extract interesting information or knowledge from the text documents. In the work, a discussion over framework of text mining with the techniques as above with their pros & cons and also applications of Text Mining is done. In addition, brief discussion of Text Mining benefits and limitations has been presented.

Ning Zhong et al, in paper [7], stated that many data mining techniques have been proposed for mining useful patterns in text documents. However, how to effectively use and update discovered patterns was still an open research issue, especially in the domain of text mining. Since most existing text mining methods adopted term-based approaches, they all suffer from the problems of polysemy and synonymy. Over the years, people have often held the hypothesis that pattern (or phrase)-based approaches should perform better than the term-based ones, but many experiments do not support the hypothesis. The paper presented an innovative and effective pattern discovery technique which includes the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves encouraging performance.

Text mining is the discovery of interesting knowledge in text documents. It was a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. In the beginning, Information Retrieval (IR) provided many term-based methods to solve the challenge, such as Rocchio and probabilistic models [8], rough set models [9], BM25 and support vector machine (SVM) [10] based filtering models.

Business Intelligence (BI) is a process for increasing the competitive advantages of a business by intelligent use of available information collection for users to make wise decision [11], [12]. It was well known that some techniques and resources such as data warehouses, multidimensional models, and ad hoc reports are related to Business Intelligence [3]. Although these techniques and resources have served us well, they do not totally cover the full scope of business intelligence [13].

In our paper, we consolidated and improve with more special features and propose our method.

III. PROBLEM STATEMENTS

Internet has significantly abridged the time taken to send a résumé by the job seekers, but the recruiter's work has become more complex because with this technological advancement they get huge volume of résumés for each job opening. It becomes practically impossible to manually scan and analyze each résumé that meets their organization's job requirement. Most of the existing approaches focus on either parsing the résumé to get information or employing some customized filtering methods to cater to their needs. Moreover, résumés differ in format and style, making it cumbersome to maintain a uniform structural repository which would contain all the necessary relevant information. Limited amount of research has been carried out on filtering the best match for a particular requirement. Recruiters have to scan all the similar looking résumés manually, after applying the filters.

IV. PROPOSED METHOD

The core objective of the paper is to develop a methodology to mine the useful information from the unstructured textual content in order to improve the business intelligence. The mining process can be achieved by new emerging technology, which is variant from data mining.

With the help of text mining, the user can able to discover previously unknown knowledge in text, by automatically extracting information from different written resources developed in natural languages. It can be now familiar because of its approaches to information management, research and analysis. Thus, text mining is the extension of data mining and obtains the goal of extracting meaningful data from different sources of textual documents.

In data mining, the collection of data is stored in the repository known as Data Warehouse. Likely, in text mining, the collection of documents is stored in the repository known as Document Warehouse. From this Document Warehouse, the text has to be extracted using text mining. The summary of the proposed methodology to extract the text from different sources of documents and make the extracted text by the decision makers to support for business intelligence is as described below:

The user may wants to mine the text from different sources of documents stored in a word file or an excel file or in any text or pdf file. We have to propose a generalized methodology in order to extract the data from any sources of files. To perform this, a configuration file is developed to ensure that it support all kinds of documents. In this configuration file, a set of configurations with suitable conditions in order to train the data are provided. This can be realized using the regular expression such as [0-9], [a-z], [A-Z]. Based on this regular expression, the text can be identified and extracted. The configuration file used to classify the document using the regular expression is given below:

4.1 Configuration File

```
<xml version="1.0">
<!—Configuration File for Text Mining -- >
<config>
<exp1>
        <cond> [0-9] </cond>
        <value> Numeric </value>
</exp1>
<exp2>
        <cond> [a-z] || [A-Z] </cond>
        <value> Name </value>
</exp2>
<exp3>
        <cond> [0-9] && count =10 </cond>
        <value> Mobile </value>
</exp3>
<exp4>
        <cond> [a-z] && [. ,] </cond>
        <value> Qualification </value>
</exp4>
<exp5>
        <cond> [@] </cond>
        <value> E-Mail Id </value>
</exp5>
<exp6>
        <cond> [0-9] </cond>
        <value> Years of Experience </value>
</exp6>
<exp7>
        <cond> [a-z]||[A-Z] &&[0-9] </cond>
        <value> Skillset </value>
</exp7>
<exp8>
        <cond> [DD-MM-YY] || [DD-MM-YYYY] </cond>
        <value> Date </value>
</exp8>
<exp8>
        <cond> [0-9] </cond>
        <value> No of jobs switched </value>
</exp8>
</config>
</xml>
```

From the above illustrated configuration file, the field and its data can be identified using the regular expression defined in <cond>.

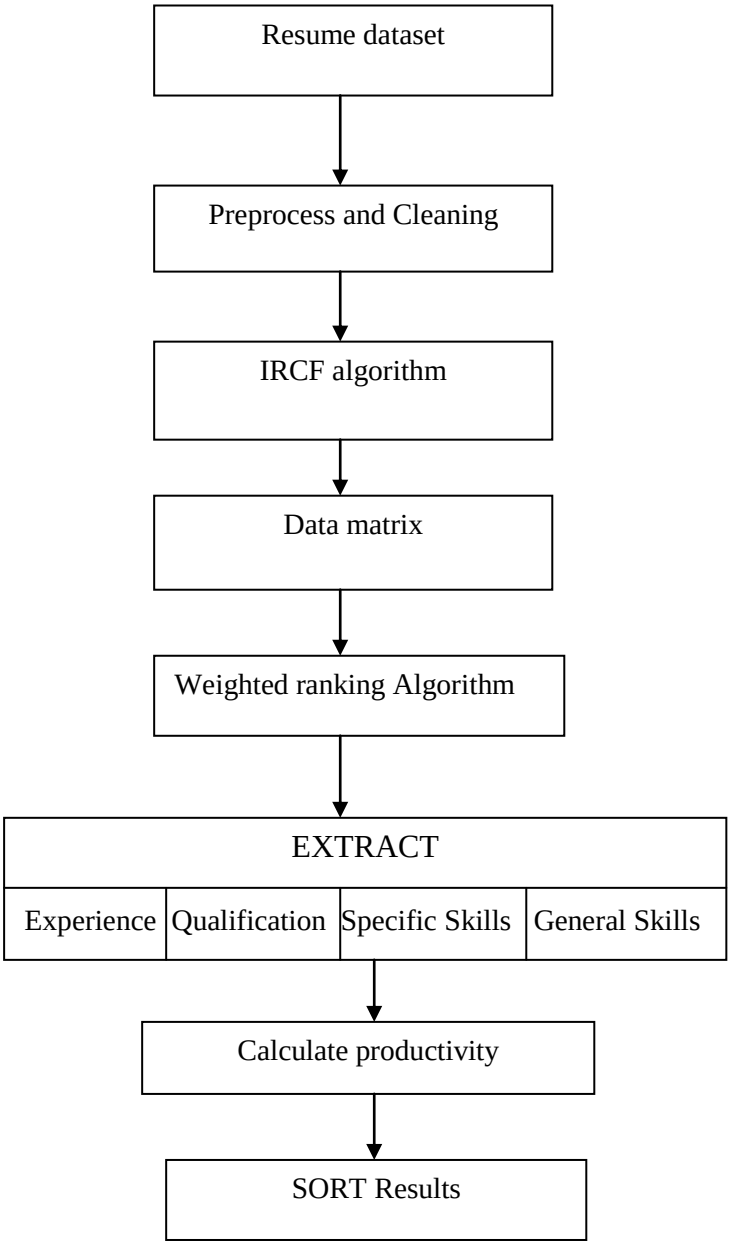


Fig. 1 Proposed Architecture

The basic steps involved in the proposed methodology to extract the text from the document based on the configuration file defined are enumerated below: The document may be of different types and each type has individual document reader.

The steps are

- Identify the document type
- Develop a Configuration File based on Document type
- Choose the Document Reader
- Categorize the text in the document

➢ Summarize the Result

The document warehouse consists of set of documents from which we have to extract the useful text by using our proposed methodology.

1. The initial step of our process is to identify the source of document from the data warehouse.

$$S = \sigma (D / N)$$

Where

S – Source of document

σ (D / N) – Selection of a document from N document in Warehouse

2. The next process is to identify the type of the chosen document, D. The type may be of word file, excel file, html file, or pdf file.
3. The next step is to develop a configuration file with set of conditions to extract the required information based on the user requirement.

Config, ζ = {x1 → x2 / x1 is the set of conditions to validate the text in the document,
x2 is the set of values for the condition x1, where x1,x2 € X,
X is the selected Document from the warehouse}

Following the creation of configuration file, the next process involved in this mining process is to choose the document reader to read the document. The document reader can be chosen based on the document type.

Based on these 3 steps, the next step is to extract the text from the document by reading the document using document reader and to identifying the text using the configuration file developed in step2.

$$\lambda = \Pi\, d1 \in D$$

$$\alpha = \lambda / \zeta$$

where,

ζ – configuration file

λ– text, d1 read by the document reader from selected document, D

α– extracted text by verifying the condition given in configuration file, ζ with λ.Finally the resultant obtained in α is summarized by storing it a 2-dimensional array. The resultant obtained may be look like as follows:

α1	Field1	Value1
α2	Field2	Value2
.		
.		
.		
αn	Fieldn	Valuen

Finally, by reading the values from this 2-dimensional array, the resultant value has to be concatenated to form the textual data. Or it has to be stored in a table in an ordered manner.

Thus, the text mining process has to be carried out successfully by extracting the useful information from the document stored in a document warehouse. From the mined result, the business executives can make the decision to improve the business intelligence.

4.2 Algorithm

Algorithm : IRCF Text Mining Algorithm

Input : Document from the Document Warehouse, (D / N)

Supported Input: Configuration File, ζ developed for D (as in section Configuration File described above)

Output : Resultant required mined text, α

```
Begin
Read the Inputted textual document, D
Analysis the Configuration File, ζ
Choose the Document Reader for D, DR
i = 1
k = 1
Repeat
  DR = DLi / D, where DLi is the read line by DR
  For j = 1 to ζ.End
    If DR.equals(ζ.condj) then
          Field = ζ.valuej
          Value = ζ.condj
          End loop
    End if
  Next
  Arr[i][k] = {DR, Field, Value}
  k++
  i++
Until DR = EOF
For a=1 to n
  For b=a+1 to n
    Navigate (Arr[a][b] → table[a][b]);  //store the array value
                                          to table
  Next
Next
End
```

4.3 Algorithm Explanation

The process of algorithm *IRCF Text Mining Algorithm* is explained in this section. (IRCF – Information Retrival based on Configuration File). The initial step is to get the input document from the user. The document can be chosen from the document warehouse. Also, the configuration file is inputted for the algorithm to extract the text from the document. From these two inputs, the mining process can be performed as follows:

First, with the help of the chosen Document Reader, the Document is read line by line. Upon fetching each line, it is passing through the configuration file to check for the condition. If the condition with the configuration file matches with the read line, then the value is fetched from the configuration file and stored in the two dimensional array. This process is iterated until the document reader reaches the end of the file.

Upon completing this process, the final step is to read the data from the two dimensional array and store the value in the corresponding field in the table. Thus the extraction process is completed successfully and the text from unstructured document is converted into a structured table. From the value stored in the table, the decision maker can make the decision to improve their process.

4.4 Weighted Ranking Algorithm

The classified data from the IRCF algorithm are populated in the two dimensional array and this data is used to rank the resumes according to the requirement after calculating the weightage of each resumes related to experience, skillsets, qualification and frequency of job switched. The weightage for the attributes are enumerated clearly in the table 4.2

Attribute	Prefixed Character	Weightage
Qualification		
Bachelors	B	1
Masters	M	2
Doctorate	D	3
Experience		
1 Years	Y	1
2 Years	Y	2
N Years	Y	N
Skills		
General Skills	G	1
Specific Skills	S	2
Age		
<= 25	A	1
>25 - 30	A	2
>30 - 40	A	3
>40	A	4

Table 4.2 – Weightage for Attributes

Algorithm – Weighted Ranking

Inputs : Dataset Ds

Supported Inputs : User Criteria Uc

Outputs : Ranked index of Resumes

Begin :

Load Dataset Ds

Calculate Weightage for Uc

Calculate Productivity Weightage Pw

Pw= (Experience Yn x S S) + (Yg x G)

Sort DESC Pw according to Qualification

Store Pw according to Qualification

Return : Ranked Index of Resumes

End:

The output obtained from the ICRF algorithm is considered as the input dataset for the Weighted ranking algorithm. The data set is loaded first and the attribute weightage is obtained from the dataset with the prefixed values enabling to find the exact criteria. The productivity Weightage Pw is calculated using the following formula

Yg= Total experience – Specific Skill Experience

Pw= (Experience Yn x S) + (Yg x G)

Where S = Specific Skills, G=General Skills

Let us consider a person with 4 years of experience in Crystal Reports 10.0 out of his 6 years of total experience, then Pw= 4

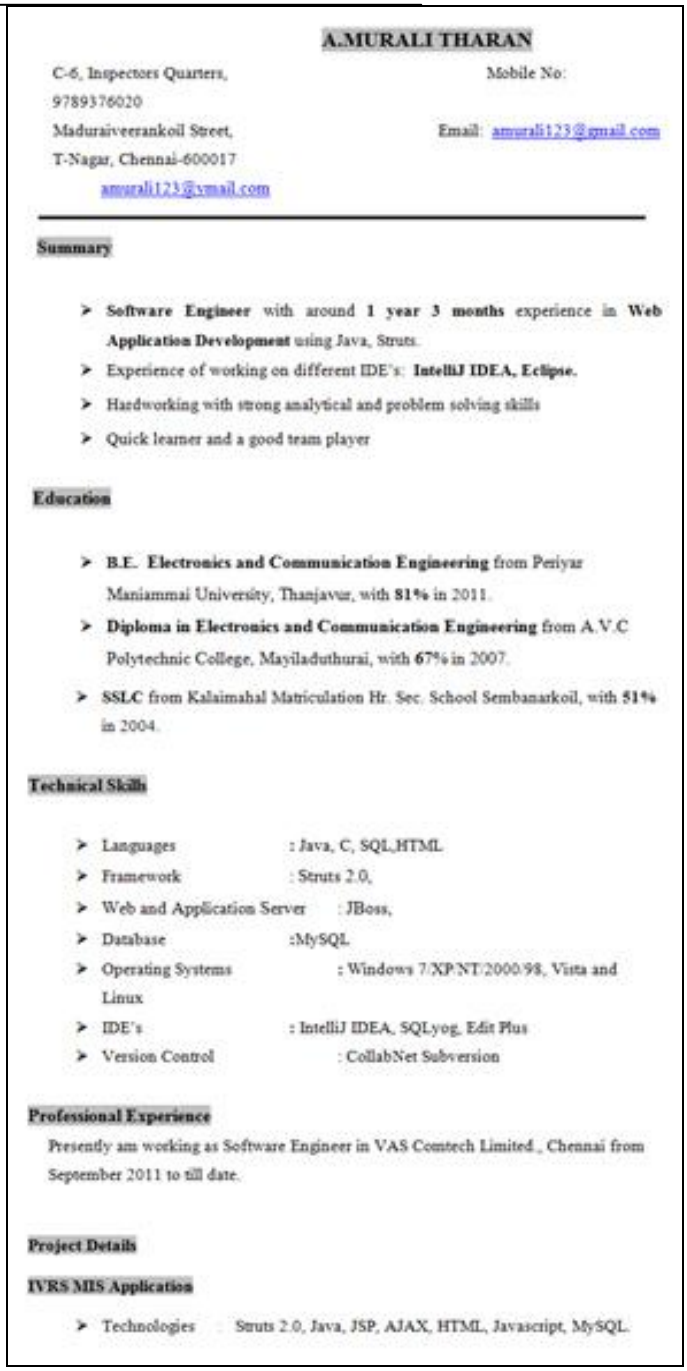
x 2 + 2 x 1=> 10 The productivity weightage calculated for this person is 10. Finally the calculated Pw is sorted descending according to the qualification weightage and stored for decision making.

V. EXPERIMENTAL RESULTS

The proposed methodology has been experimentally verified and the result proves that the proposed methodology satisfies the aim of the paper. To do the experiment, we have undertaken 300 candidate's resume, which is of rtf format. From those 300 candidates, a particular group of members has to be selected. First, the selection process can be carried out using the existing algorithm namely "*KNN Text Classification Algorithm"* by examining each resume and identifying the qualifying candidate. The execution time and the scan time are noted and compared with that of the proposed algorithm.

Next, the selection process can be carried out by using the proposed IRCF text mining algorithm. Based on the algorithm, the qualifying candidate can be chosen and the time period to obtain the result has to be noted. By comparing these two time periods, we can conclude that the proposed algorithm performs well than the existing method. Thus our proposed methodology satisfies the aim of the paper and helps the decision makers to identify and recruit deserving candidates with the skillsets and the criteria to match with that of the company policies during recruitment process.

5.1 Sample Resume



A.MURALI THARAN

C-6, Inspectors Quarters, Mobile No: 9789376020
Maduraiveerankoil Street, Email: amurali123@gmail.com
T-Nagar, Chennai-600017
amurali123@ymail.com

Summary

- **Software Engineer** with around **1 year 3 months** experience in **Web Application Development** using Java, Struts.
- Experience of working on different IDE's: **IntelliJ IDEA, Eclipse.**
- Hardworking with strong analytical and problem solving skills
- Quick learner and a good team player

Education

- **B.E. Electronics and Communication Engineering** from Periyar Maniammai University, Thanjavur, with **81%** in 2011.
- **Diploma in Electronics and Communication Engineering** from A.V.C Polytechnic College, Mayiladuthurai, with **67%** in 2007.
- **SSLC** from Kalaimahal Matriculation Hr. Sec. School Sembanarkoil, with **51%** in 2004.

Technical Skills

- Languages : Java, C, SQL,HTML
- Framework : Struts 2.0,
- Web and Application Server : JBoss,
- Database :MySQL
- Operating Systems : Windows 7/XP/NT/2000/98, Vista and Linux
- IDE's : IntelliJ IDEA, SQLyog, Edit Plus
- Version Control : CollabNet Subversion

Professional Experience

Presently am working as Software Engineer in VAS Comtech Limited., Chennai from September 2011 to till date.

Project Details

IVRS MIS Application

- Technologies : Struts 2.0, Java, JSP, AJAX, HTML, Javascript, MySQL.

International Journal of Advance Research in Science and Engineering
Vol. No.4, Special Issue (01), November 2015
www.ijarse.com
IJARSE
ISSN 2319 - 8354

VI. FINDINGS AND ANALYSIS

No. of Resumes	Execution Time in (sec)	
	KNN	IRCF
50	82	43
100	157	78
200	340	168
250	504	248
300	712	341

Table 6.1 – Comparison Data

The Proposed algorithm IRCF is compared with KNN algorithm and from the above table 6.1 it is obvious that the execution time of IRCF is low when compared to KNN.

The Resume Relevancy ratio is first found after obtaining the weighted ranking resume set.

$$\text{RR ratio} = \frac{\text{Total Number of relevant resumes}}{\text{Total Number of resumes in Dataset}}$$

The actual Resume Relevancy ratio is calculated manually using the following formula

$$\text{ARR ratio} = \frac{\text{Total Number of actual relevant resumes}}{\text{Total Number of resumes in Dataset}}$$

The algorithm is tested to find the probable candidates with master's degree, minimum of 3 years of experience in crystal reports 10.0 and overall good experience in reporting tools and Microsoft technologies. The proposed algorithm fetched 6 resumes out of 300 with productivity weightage ranging from 11 to 17.

Ratio	Condition 1	Condition 2	Condition3
RR ratio	0.671	0.88	0.656
ARR ratio	0.7	0.71	0.68

Table 6.2 – Ratio Calculation Result

VII. CONCLUSION

The proposed algorithm IRCF with weighted Ranking methodology performs well in extracting relevant resumes from the unstructured document without any manual intervention. The proposed algorithm ICRF clearly outperforms the KNN algorithm in terms of execution time.

REFERENCES

[1] Atika Mustafa, Ali Akbar, and Ahmer Sultan, "Knowledge Discovery using Text Mining: A Programmable Implementation on Information Extraction and Categorization", International Journal of Multimedia and Ubiquitous Engineering, Vol. 4, No. 2, April, 2009.

[2] Garcia, Cristina Cristina Bicharra, "The Role of Domain Ontology in Text Mining Applications: The ADDMiner Project", Data Mining Workshops, International Conference on Dec. 2006.

[3] Bhujade, Vaishali, "Knowledge Discovery in Text Mining Technique using Association Rules Extraction", Computational Intelligence and Communication Networks (CICN), International Conference on oct. 2011.

[4] Raymond J.Mooney and Un Yong Nahm, "Text Mining with Information Extraction", Proceedings of the 4th International Conference, 2005.

[5] Li Gao, Elizabeth Chang, and Song Han, "Powerful Tool to Expand Business Intelligence: Text Mining", Engineering and Technology, 2007.

[6] Divya Nasa, "Text Mining Techniques – A Survey", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012.

[7] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, "Effective Pattern Discovery for Text Mining", IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, January 2012.

[8] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.

[9] Y. Li, C. Zhang, and J.R. Swan, "An Information Filtering Model on the Web and Its Application in Jobagent," Knowledge-Based Systems, vol. 13, no. 5, pp. 285-296, 2000.

[10] S. Robertson and I. Soboroff, "The Trec 2002 Filtering Track Report," TREC, 2002, trec.nist.gov/pubs/trec11/papers/OVER. FILTERING.ps.gz.

[11] B. de Ville, "Microsoft Data Mining: Integrated Business Intelligence for e-Commerce and Knowledge Management", Boston: Digital Press, 2001.

[12] P. Bergeron, C. A. Hiller, "Competitive intelligence", in B. Cronin, Annual Review of Information Science and Technology, Medford, N.J.: Information Today, vol. 36, chapter 8, 2002.

[13] M. J. A. Berry, G. Linoff, "Data Mining Techniques: For Marketing,Sales, and Customer Relationship Management", Wiley Computer Publishing, 2nd edition, 2004.