



Resume Information Extraction using Feature Extraction Model

V. Jayaraj¹, V. Mahalakshmi², P. Rajadurai³

¹Associate Professor, ^{2,3}Research Scholar

^{1,2,3}Department of Computer Science and Engineering,

Bharathidasan University, Tiruchirappalli – 620023, Tamil Nadu, India

Abstract

Background: A novel algorithm named feature extraction is used to extract the textual data.

Method: The most common method used for feature extraction from the documents is TF-IDF (Term Frequency- Inverse Document Frequency). The TF-IDF measure is a method much used for weighting terms in information retrieval.

Results: The basic idea of this research work is to develop an approach to select the appropriate resume efficiently and enhances the recruitment process by extracting the unique and special features in the resume and makes it simpler for the employer to select the right candidates without much effort and manual work.

Keywords: Term Frequency, Inverse Document Frequency, Patterns, Unstructured data.

I. Introduction

The last few decades has witnessed a stupendous growth of information across the internet. The giant of information are unused across the globe and it requires rigid methodology to mine and extract the text. The growth of information is increasing exponentially and it becomes more important to detect useful pattern from the data. It is very difficult for the user to retrieve the data from the database. To solve this problem many techniques have been implemented and still require enhancements to overcome many retrieval problems in the unstructured data (Jayaraj *et al.*, 2014). Text mining is a process to empathize and discover useful meaningful tacit information from a large amount of the semi-structured or unstructured textual data.

Simply the text mining comprehends the intermingling of human linguistic competence and computational power of the system (Fan *et al.*, 2006). The linguistic capability includes the adeptness to differentiate spelling, filtering out unpromising data, understanding the synonyms/meaning, different slangs, abbreviations and finding the literal meaning. Orthodox approaches in text clustering and mining use words as a measure to discover similarity between documents. These words are presumed to be reciprocally independent which in real application it may differ and the concept, semantics and features are what describe the documents. The technique of extracting these features from the documents is called feature extraction (Liu *et al.*, 2005). The concept of feature extraction has been successfully practiced in unsupervised algorithms like PCA (Principal Component Analysis) and SVD (Singular Value Decomposition). Recently most research aimed to speed up text mining process involves improvements in extracting features from the text, since the time consumed for extracting the word features from texts surpasses the initial training time.

This paper portrays a fast method for the extraction of features with the aid of a configuration file to figure out the unpromising texts and completely eliminate the texts to reduce the dimensionality and space (Dorre *et al.*, 1999). The most important advantage of this work is after extracting the promising features from the text, the feature selection process is carried out to filter out the unwanted meaningless text from the textual data. This new approach reduces the space and the dimension of the text considerably. The feature extraction phase is further subdivided into two levels, namely extraction and selection. The extraction reduces the space considerably and when further selection is carried out the space is reduced largely. But the feature extraction phase involves large complexities and limitations. Extraction of the information from resumes has been an important area of focus for a lot of researchers. A resume is a concise document about an individual trying to market him/her to the industry. Resumes contain both structured and unstructured data too (Kun Yu *et al.*, 2005). Most of the business records are maintained in the form of documents and hence the documents are in unstructured format (Jayaraj *et al.*, 2015). In a resume, the format is not predetermined and it is based on the authors thinking and writing skills, which makes the information extraction, comparison, and selection a

intimidating task. Each resume is unique in its own way as it contains words and sentences as features (Sumit Maheswari *et al.*, 2010). It can also be viewed as a multi-section document, with description of each section, which highlights the different aspects of an individual's professional career (Abhishek Sainani, 2011). A recent survey found that about 78% of the employers were content with investing in the online recruiting domain (Hawkins *et al.*, 2004). Every resumes are scanned and processed using information extraction methods, which are then stored in the employers resume repository (Xing Yi *et al.*, 2007). When a recruiter needs some resumes, he/she performs a keyword search on this resume database. This search uses dynamic filtering techniques. The set of filters help in tossing out irrelevant resumes. Google uses "resume filters" to separate through the piles of resumes it gets every month for every job posted. Experts and recruiters state that using resume filters purges around 60 to 75% of the unwanted documents. However, applying strict filters to a set might unfavourably affect the search (Norm Schneider *et al.*, 2011). Enforcing the search criteria may result in certain prospective candidates resumes being ignored. Even though strict filters might neglect certain potential resumes, the speed and accuracy with which the results are computed are so imposing that the recruiters and hiring managers have started to favor electronic methods for determining suitable candidates (Abhishek Sainani, 2011).

II. Existing Feature Extraction Approaches

Generally various techniques are employed in extracting features from the textual data and some of the commonly used approaches are illustrated in this section. (Loic Lecerf *et al.*, 2009) came up with an algorithm for extracting features from the documents which works in a recursive manner and used the following patterns for extraction Separator Set, Factual Patterns, Generalised patterns and Compound patterns. The frequent pattern generator calls recursively Algorithm on each basic pattern $p \in P$. It profits from the regular pattern match on arbitrary strings, according to which $\text{match}(p_1, d) = (a, b)$ and $\text{match}(p_2, d) = (b + 1, c)$ implies $\text{match}(\text{concat}(p_1, p_2), d) = (a, c)$ where d is an input document. The algorithm tries to compound a new pattern by checking whether the ending position of a current pattern is followed by the beginning position of a basic one.

FeatureExtractAlgo

Require: cur -current pattern, $Mc = \{(b_i, e_i)\}$ - the match set of basic patterns

Ensure: RES - all frequent patterns prefixed with cur

1. $RES := \emptyset$
2. for all $p \in P$ do
3. Let $Mp = \{(b_j, e_j)\}$ be the match set of p
4. $M := \{(b_i, e_j) | e_i + 1 = b_j, e_i \in Mc, b_j \in Mp\}$
5. if $|M| > k$ then
6. $new := \text{Concat}(cur, p)$
7. $RES := RES \cup new \cup \text{FeatureExtractAlgo}(new, M)$
8. End if
9. End for
10. Return RES

Figure 1: FeatureExtractAlgo Pseudo code

The accuracy of this algorithm ranges between 65 % and 76 %. This algorithm works using regular expressions and the email extracted from the text is shown here under

1. $Re = "\.com\b":$ is 1 if the text contains the '.com' substring.
2. $Re = "@[a-z]\{3,5\}":$ is 1 if a text contains '@' followed by 3 to 5 characters in the a - z range.
3. $Re = "\b\w\.\w\b":$ is 1 in the presence of two alphanumeric strings separated by a dot.

This is how the email text is extracted from the textual data automatically.

III. Term Frequency- Inverse Document Frequency

In Information Retrieval, this process is sometimes called index term selection. The reason for performing feature extraction is that using the set of all words in a document collection to represent its documents generates too much noise (Baeza *et al.*, 1999). This is especially important when mining for association rules, where the user must be able to make sense of the rules in the result set, and the computational cost of using all words is very high. The most common method used for feature extraction from the documents is TF-IDF (Jayaraj *et al.*, 2015). The TF-IDF (Term Frequency- Inverse Document Frequency) measure is a method much used for weighting terms in information retrieval. It can be used both to find the potentially most meaningful document features among the native document features in a document, or to filter the document features found by using another method. TF-IDF works by assigning a weight to each term, or document feature, in the document collection. Then, either the n highest ranked terms or all terms above some minimum weight threshold may be

kept. The TF-IDF measure consists of two separate measures. The TF part is the occurrence frequency of a term inside a document, while the IDF part is the occurrence of inverse frequency of a term among all documents in the collection. The last part thus reduces the weight of terms appearing in many documents.

Let N be the number of documents.

n_i = number of documents where term k_i occurs

$\text{freq}_{i,j}$ = the frequency of term k_i in the document d_j

$\text{max}_i \text{freq}_{i,j}$ = number of times the most frequent term present in document d_j

The normalized frequency is given by

$$\text{TF}_{i,j} = \frac{\text{Freq}_{i,j}}{\text{Max}_i \text{freq}_{i,j}}$$

The inverse document frequency for term k_i is given by

$$\text{IDF}_{i,j} = \log \frac{N}{n_i}$$

The weight of term k_i is given by

$$W_{i,j} = \text{TF}_{i,j} \times \text{IDF}_{i,j}$$

Gladis Puspha Rathi *et al.*, (2012) used one of the common feature extraction techniques to extract the meaningful texts from the documents. The feature extraction technique is employed for extraction and the extracted features are applied to support vector machines to obtain higher accuracy in classification.

IV. Proposed Feature Extraction Technique

The dataset used in this paper is quite different from the contemporary textual dataset and a customized feature extraction technique is imperative to extract promising texts and to reduce the dimension of the data. Here few promising features like emails, age, general skillset, specific skillset, experience, qualification of the candidates are extracted using a configuration file. The difficulty in extracting this important textual information needs utmost care and the process of extraction is showcased here under.

The Email ID plays a pivotal role enhancing the productivity of the recruitment process in an organization and hence this textual information needs to be extracted from the resume data. The email ID are extracted from the data using regular expressions, like

Expr1 = "\.com\b": is 1 if the text contains the '.com' substring.

Expr2 = "\.in\b": is 1 if the text contains the '.in' substring.

Expr3 = "\.net\b": is 1 if the text contains the '.net' substring.

Expr4 = "\.biz\b": is 1 if the text contains the '.biz' substring.

Expr5 = "\.org\b": is 1 if the text contains the '.org' substring.

Expr6 = "[a-z]{3,5}@": is 1 if a text contains '@' followed by 3 to 5 characters in the a - z range. Sample Resume with email ID in table 1 is extracted from sample resume datasets.

Here care should be taken to trim the word email from the textual content and this is carried out by the configuration file and the extracted email is stored in the two dimensional array with the field name and value.

Resume ID	Email ID value
R43	manoj.pm2011@gmail.com
R187	Koushik.vlb@gmail.com
R3872	b.rutuparna04@gmail.com

Table 1: Extracted Email ID from sample resume datasets

Most of the resume does not have a field named age instead date of birth or DOB field will be present. if the age value is directly provided in the resume the process is straight forward and the value is retrieved and stored. If the age field is not provided then the algorithm will check for date of birth or DOB. If date of birth text is texted in the resume, a small calculation is needed to convert the date of birth to age. Here again regular expressions are used to retrieve the DOB for calculation. Here again a complexity arises due to the format of the date, since every individual uses different formats like "dd/mm/yyyy", "dd/mm/yy", "dd/MON/yyyy", "dd MON yyyy". Here care should be taken during extraction.

Expr1 = “\date of birth\ [0-9]{6-8}[/]{2}” is 1 if the text, date of birth followed by 6 to 8 numbers and two “/”.
 Expr2 = “\DOB\ [0-9]{6-8}[/]{2}” is 1 if the text, DOB followed by 6 to 8 numbers and two “/”.
 Expr3 = “\date of birth\ [0-9]{4-6}[a-z]” is 1 if the text, date of birth followed by 4 to 6 numbers and a-z characters.

Expr4 = “\DOB\ [0-9]{4-6}[a-z]” is 1 if the text, DOB followed by 4 to 6 numbers and a-z characters.

From the extracted date of birth, the year is fetched separately and that year value is subtracted from the current year. For example, if the extracted date of birth is 13/09/1980. The Year 1980 is fetched from the DOB and it is subtracted from the current year 2015. Age = 2015 – 1980 = 35 Years

Some of the resume will neither have date of birth nor age. In such cases the algorithm checks for the qualification or the year of passing the school education. Generally in India, the candidate will pass the secondary school at the age of 15. This 15 year can be subtracted from the year of passing the secondary school to find the exact age. By using this dataset we also fetched many results like qualification, experience, and specific skillset.

V. Mining Qualification Information from Resume Dataset

The extraction of qualification is not as tedious as age extraction since the qualification will be a straight forward one. From the configuration file, the section named academic, qualification or professional qualification will aid the qualification extraction from the resume dataset. The extracted qualification is processed to find the weightage and the qualification weightage is stored for easy and faster clustering process and shown in figure 1. Usually the highest qualification is considered by the employers and here the highest qualification is extracted to calculate the weightage. In this paper more than 10000 resumes are used for extracting information from resumes

Academic Profile

Course	Institute	Year of Passing	Percentage
MCA	Nallamuthu Goundar Mahalingam College, Pollachi.	2008	79%
B.Sc(CS)	Sree Saraswathi Thyagaraja College, Pollachi.	2005	65.7%
HSC	P.G.H.S.School, Pollachi.	2002	65.6%
SSLC	C.H.S.School, Pollachi	2000	87.2%

Figure 1: Highest Qualification extraction from Sample resume ID 310

Resume ID	Weightage	Qualification	Education weight
310	2	M2	50
28	2	M2	50
287	2	M2	50
236	2	B2	30

Table 2: Extracted and calculated Qualification

The qualification information extracted are calculated for weightage and stored for further processing as shown in the table 2. The extracted features or interesting and promising texts are then converted into predefined text according to the attribute and weightage for smoother and faster clustering process.

VI. Conclusion

This work has made an extensive effort to extract the special features present in the resumes. The basic idea of this research work is to develop an approach to select the appropriate resume efficiently and enhances the recruitment process by extracting the unique and special features in the resume and makes it simpler for the

employer to select the right candidates without much effort and manual work. It also shows that applying different IE models in different passes according to the contextual structure is effective). The important features present in the resumes are extracted and the principle behind the extraction is clearly showcased in this paper.

References

- [1] Abhishek Sainani, "Extracting Special Information to Improve the efficiency of Resume Selection Process," June 2011.
- [2] B. L. Hawkins, J. A. Rudy and W. H. Wallace "Recruiting, Retaining, and Reskilling Campus IT Professionals. Technology Everywhere: A Campus Agenda for Educating and Managing Workers in the Digital Age". Dolan, A. F. 2004. Jossey-Bass: 75--91.
- [3] Baeza Yates, Ricardo, and Berthier Ribeiro-Neto. Modern information retrieval (1999). New York: ACM press, 463.
- [4] Dorre, Jochen, Peter Gerstl, and Roland Seiffert (1999). Text mining: finding nuggets in mountains of textual data. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 398-401. ACM, 1999.
- [5] Fan, Weiguo, Linda Wallace, Stephanie Rich, and Zhongju Zhang (2006). Tapping the power of text mining. *Communications of the ACM*, 49(9): 76-82.
- [6] Jayaraj, V., and V. Mahalakshmi. "Augmenting Efficiency of Recruitment Process using IRCF text mining Algorithm." *Indian Journal of Science and Technology* 8.16 (2015).
- [7] Jayaraj, V., and V. Mahalakshmi. "Information Retrieval Configuration File Text Categorization Algorithm for Improving Business Intelligence." *International Journal Of Computational Engineering And Management* (IJCEM), ISSN:2230-7893, January 2015.
- [8] Kun Yu, Gang Guan, and Ming Zhou, "Resume information extraction with cascaded hybrid model". In *ACL '05: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 499–506, Morristown, NJ, USA, 2005.
- [9] Lecerf, Loic, and Boris Chidlovskii (2009). Scalable Feature Extraction from Noisy Documents. *Document Analysis and Recognition, ICDAR'09. 10th International Conference on. IEEE*.
- [10] Liu, Luying, Jianchu Kang, Jing Yu, and Zhongliang Wang. A comparative study on unsupervised feature selection methods for text clustering (2005). In *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference*, 597-601.
- [11] Norm Schneider, "Job Hunters: Resume Filters May Help or Hinder Your Job Search," Published on August 23, 2011
Employment <http://bizcovering.com/employment/job-hunters-resume-filters-may-help-or-hinder-your-job-search/#ixzz1fajVNEJ2>.
- [12] Rathi, VP Gladis Pushpa, and S. Palani (2012). A novel approach for feature extraction and Selection on mri images for brain tumor Classification. *CCSEA, SEA, CLOUD, DKMP, CS & IT* 5, 225-234.
- [13] Sumit Maheshwari, Abhishek Sainani, and P. Krishna Reddy, "An Approach to Extract Special Skills to Improve the Performance of Resume Selection". In *Proceedings of DNIS. 2010*, 256-273.
- [14] V.Jayaraj and V.Mahalakshmi, "Improving Text Categorization Using Configuration File" *American International Journal of Research in Science, Technology, Engineering & Mathematics* September 2014.
- [15] Xing Yi, James Allan and W. Bruce Croft, "Matching Resumes and Jobs Based on Relevance Models". In *Proceedings, SIGIR 2007*.