ISSN:0975-9646

Talent Management in Organizations Using Mining Techniques

Manogna.N[1] , *Sumedha Mehta*[2]

[1,2]*RAC, DRDO,*
New Delhi, INDIA.

***Abstract*-An organization's resume repositories of both the external applicants and its own employees hold valuable information, which when analysed, can be used for recognizing the talent most suitable for it, as well as provide insights to the career growth and training prospects of the personnel. In this highly competitive environment it is imperative to efficiently mine this information. Several advanced computing technologies like Text Mining, Data Mining and Information Retrieval can extract high quality information pertinent to the talent acquisition and talent management in organizations. This paper provides a review of these technologies and also emphasises that an organization should develop such resume processing tools for improving its Human Resources (HR) processes.**

***Keywords*—Information Retrieval, Information Extraction, Text Mining, Data Mining**

I. INTRODUCTION

Repositories of employee resumes and candidate resumes are the vital reservoirs of information which is instrumental for HR Management executives to mine the best talents, train them, assess their performance for promotion, and ultimately keep them contented in the organization. A typical resume embodies the information about an individual's personal details, educational qualifications, experience, training, publications, and awards and so on, in semi structured or free-form text created in a multitude of document formats (text, PDF, HTML, DOC etc.). Manual analysis of the resume repository is not only costly, but also lacks objectivity and is subject to limitations in its quality depending on the expertise of the person doing it.

Automated resume processing systems are increasingly being looked upon as the solution to this problem. However, in these systems, the unstructured nature of the resumes poses a challenge to achieving high precision (fraction of retrieved instances that are relevant) and recall (fraction of relevant instances that are retrieved) while finding, for instance, a resume matching a particular job specification. The resumes contain no fixed set of sections, no fixed ways of indicating section headers, may contain tables and images, multiple columns, domain-specific vocabulary, fragmentary text with abbreviations and so on because of which information extraction becomes difficult. There are no readily available fixed dictionaries for the entities like skills and designations. Also, the way in which the factual information (for example, "Jan. 31, 2008" or "31-Jan-2008", marks written as % or grade points) is used, differs from one to another.

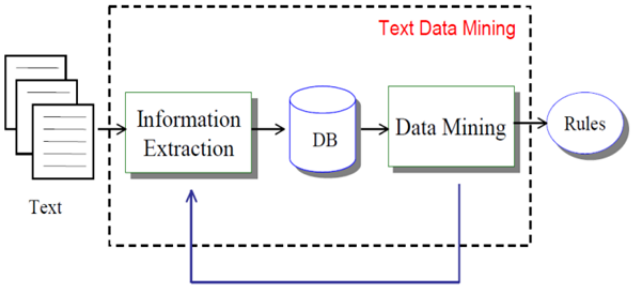


Figure1. Abstract Model of Information Extraction based Text Mining [1]

Some existing solutions for resume parsing extract some simple kinds of structured information and store it in a structured data repository. Many of the resume mining techniques involve the use of natural-language information extraction**.** The problem of structuring the information to support its processing is addressed by annotating the data with semantic mark ups using natural language processing systems such as the General Architecture for Text Engineering (GATE). Figure 1 illustrates how Information Extraction plays an obvious role in distilling structured data from unstructured text by identifying references to named entities as well as stated relationships between such entities which can then be further analyzed with traditional data-mining techniques to discover more general patterns. Moreover, the predictive rules acquired by applying data mining techniques can further be fed back to the learned information extractor to improve its recall.

The amalgamation of data and text mining is referred to as "Duo-mining" [2]. In addition, Information Retrieval techniques generally using the "Bag-of-words" model [3] can be employed for document matching and ranking. In this paper, the various approaches of mining resumes are reviewed. The paper also reviews how organizations can effectively manage their talent using these mining techniques.

II. A POTPOURRI OF RESUME MINING TECHNIQUES

We briefly review a range of technologies available to facilitate the processing of resumes for mining relevant information.

A. INFORMATION RETRIEVAL

Information Retrieval (IR) is the process of finding information resources (usually documents) of an unstructured nature that satisfies an information need from within large collections (usually stored on computers). Searches can be based on metadata or on full-text (or other

content-based) indexing. A typical document based IR system (Figure 2) consists of four main subsystems: document representation, representation of users' requirements (queries), algorithms used to match user requirements with document representations and a ranking capability. A search engine is the practical application of information retrieval techniques to large scale text collections. In a search query, Boolean logic helps defining the logical relationship between multiple search terms.

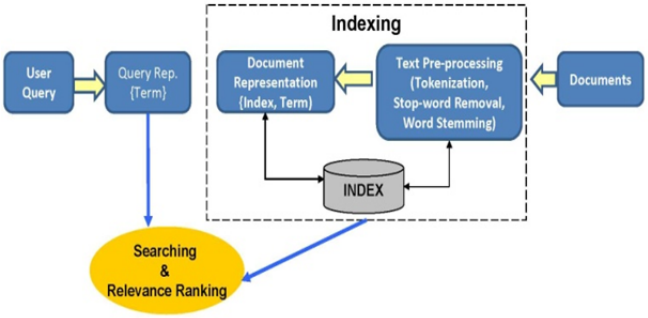


Figure 2. A graphic representation of the IR process

Document Pre-processing involves tokenizing the document stream into desired retrievable units (tokens). Usually a token is defined as an alpha-numeric string that occurs between white space and/ or punctuation. The document is then processed for deletion of stop words which help conserve precious system resources by eliminating them for further processing. A stop word list typically consists of those words known to convey little substantive meaning, such as articles (*a, the*), conjunctions (*and, but) etc.* To remove the stop words, an algorithm compares the potential index terms in the documents against a stop word list and eliminates the matched words from inclusion in the index for searching. Finally the document pre-processor removes word suffixes, commonly known as stemming. Stemming improves the efficiency of the system by reducing the number of unique terms in the index. It also improves the recall as all documents including words in various forms of the stemmed base word have equal likelihood of being retrieved. On completion of document pre-processing, the remaining entries stored on an inverted file that lists the index terms, their location and frequency of occurrence.

Searching and Retrieving these documents, is the third step in the process. IR can be performed with a vector space model, in which vectors are used to represent word frequencies. Vector-space models rely on the premise that the meaning of a document can be derived from the document's constituent terms. They represent documents as vectors of terms $d=(t_1,t_2,\ldots,t_n)$ where t_i $(1<=i<=n)$ is a non-negative value denoting the single or multiple occurrences of term i in document d. Hence, each term t of the dictionary is considered as a dimension. A document d can represented by the weight of each term: V (d) = ($w(t_1, d)$, $w(t_2, d)$, . . ., $w(t_n, d)$). One of the optimal schemes for term weight assignment is TF/IDF (Term Frequency/ Inverse Document Frequency) weighting. This algorithm measures the frequency of occurrence of each term within a document (TF). Then it compares that frequency against the frequency of occurrence in the entire database (IDF). The TF/IDF weighting scheme assigns higher weights to those terms that really distinguish one document from the others.

TF(t) = (Frequency of term t in a document) / (Total number of terms in the document)

IDF(t) = $\log_e$ (Total number of documents) / Number of documents with term t in it)

The TF/IDF values can now be used to create vector representations of documents. Similarly, a query from the user is assigned a vector. This vector is then compared to the vectors of all documents in the collection. The cosine of the angle between the document vector and query vector is calculated by using the normalized dot product of the two vectors.

Ranking of documents on the basis of estimated relevance to a query is critical. A higher cosine function value denotes a closer match between the query and document. Finally, the similarity values between query and the retrieved documents are used to rank the results.

Despite the fact that such searches can be easily and efficiently performed, they suffer from some limitations. To cite an example, a search engine lacks understanding about domain concepts and relationships like “SPSS and R are both statistical analysis software packages”. Thus a search for SPSS analysts will not return resumes of R analysts. Also, search engines are unable to comprehend customized searches like “Doctorate with 5 years experience in the field of R&D Management out of which at least 2 years are in a research laboratory”. This information may not be explicitly stated in the resume but can be inferred by examining the employment record. Finally, information retrieval systems are also constrained by their limited understanding of natural languages and the semantics of entities. Hence, it will not know that “Software Quality Assurer” and “Software Tester” mean the same while searching for resumes of people having experience in “Quality Assurance”. As a result, the burden of designing complex search conditions is thrown on the ultimate users of the system. This urges the need for a universal solution to the search mechanism which is capable of interpreting the crux of the search requirement.

B. INFORMATION EXTRACTION

Information Extraction is the technology based on analysing natural language in machine readable unstructured documents in order to extract snippets of information. Automatic annotation and content extraction of multimedia documents (images/ audio/ video) are popular IE applications. The Google Search engine uses IE to extract information for its indexing. Manual regular expressions and Machine learning methods are the standard approaches for IE.

Hand-written regular expressions(RE) defining the extraction patterns used to reliably identify entities and relations within natural language text, is the lesser used method of IE due to its inherent limitations. The lexical patterns cannot exhaustively cover all types of forms and contexts in which the desired information can appear.

Repeatedly enhancing the RE to accommodate more such forms is a tedious process.

Machine learning methods trained on human annotated corpora to automatically create extraction patterns are the most widely used methods for developing potent IE systems. Some of these methods, among others are Decision Trees, Wrapper Induction and Artificial Neural Networks (ANN).

Decision Trees describe the structure of a learned function or its corresponding learning algorithm. A decision tree (Figure 3) is a simple hierarchical representation in which the initial and intermediate nodes correspond to object attributes and the terminal nodes correspond to the identities of the objects. Attribute values for an object determine a path to a leaf node in the tree which contains the object's identification.

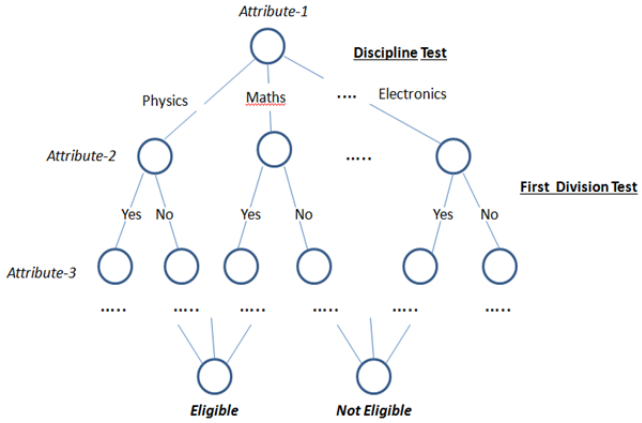


Figure 3. Segment of a Decision Tree for deciding eligibility of candidate

The knowledge base, which is the decision tree for an identification system can be constructed with a learning module. A set of the most distinguishing attributes for the class of objects being identified should be selected. As the system gains experience, the values associated with the branches can be modified for more accurate results.

Wrapper induction: Aimed at extracting information from structured and semi-structured documents on the web, a wrapper is a program code which applies a set of extraction rules to such documents and converts them into a relational form. Given a training data set, wrapper induction is a technique for automatically learning the wrappers for extracting the required information. Figure 4 describes a simple model of information extraction with wrapper induction.

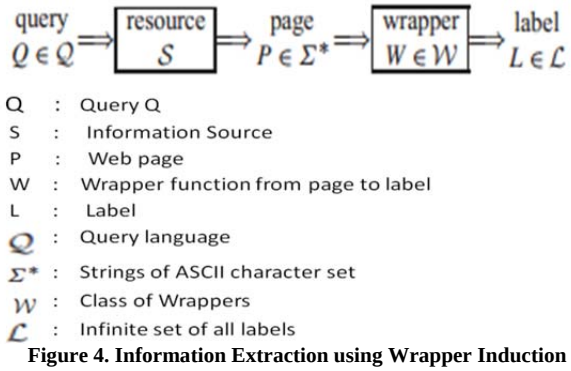
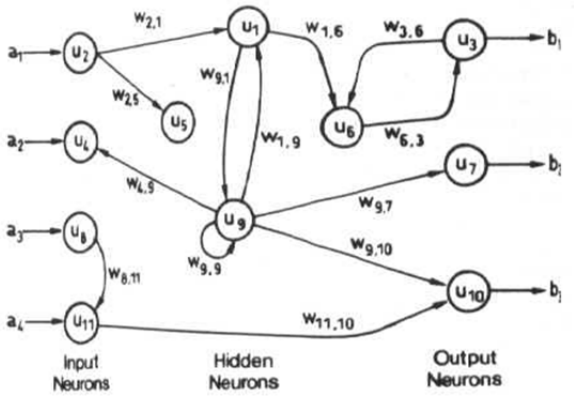


Figure 4. Information Extraction using Wrapper Induction [4]

A page's content is represented by labels, a label being a set of indices or positions in the page. The input to the learning system would be a sample of S's pages and their associated labels, and the output should be a wrapper w belonging to class W.

Artificial Neural Networks (ANN) are large networks of simple processing elements (nodes) which process information dynamically in response to external inputs. The knowledge in a ANN is in the form of weighted inter node connections ($w_{i,j}$) which form the inputs to the nodes (Figure 5). Learning the weights is performed by repeatedly presenting the network with an input pattern and a desired output response. The weights are adjusted until the difference (D) between the output response and desired response is zero.



a = Input pattern

b'= Desired output response

b= Actual response

D= b – b'

$$\mathbf{W_{new} = W_{old} + r * D * a * \gamma\, |a|^2}$$

where $0<r<1$, signifies rate of learning

Figure 5. A Multi layer Neural Network

IE systems being knowledge intensive, are difficult to build and are to varying degrees tied to particular domains and scenarios. In the present context, IE is more efficient than IR because of the possibility of dramatically reducing the amount of time people spend reading huge volumes of resumes.

C. TEXT MINING AND DATA MINING

Text Mining is the process of automatically extracting information from text based files by way of transposing its constituent words and phrases into numerical values for further analysis with data mining techniques. Data mining (or knowledge discovery from databases) is the process of extracting previously unknown, interesting patterns and models from data. Some of the text and data mining techniques discussed below can be applied to the resume documents for achieving HR business goals.

i. *Classification* is a classic data mining technique based on machine learning**.** It aims on classifying each item in a data set into one of the predefined classes. Classification method makes use of mathematical

techniques such as decision trees, neural networks etc. The classification technique is also applicable in the text domain, where the objects to be classes can be of different granularities such as documents, paragraphs, sentences or terms. for eg. classification of resume documents as "eligible" and "not eligible" by applying a classification model derived by learning a training set of resumes.

ii. *Clustering* is a data mining technique that helps construct meaningful partitions of a large set of objects by way of grouping similar objects automatically. It strives to maximize the intra class similarity and minimize the inter class similarity by making use of similarity functions. In contrast to classification, there are no predefined classes or training data. As in classification, it can also be applied in text domain for organizing documents. Clustering documents into coherent categories improves navigation and browsing in a document collection.

iii. *Regression analysis* is a data mining tool that discovers a relationship between the dependent and independent variables for predicting the future values of the dependent variable. It is used for modelling data by way of fitting an equation to a historical data set.

Text Mining using both information extraction techniques and data mining techniques on the facts generated by the information extraction phase, helps to create job market intelligence ie. the illustration of trends and patterns of the available talent pool. These trends can provide valuable guidance for optimal recruitment planning and decision making.

III. TALENT ACQUISITION AND MANAGEMENT IN ORGANIZATIONS

So far the various techniques which can be employed for mining resumes were discussed. Now we will demonstrate their application in the context of an organization.

The resume information is needed in the following stages of talent acquisition- (a) Initial screening of candidates through validation checks based on the laid down eligibility rules, (b) Short-listing of candidates for specific posts matching the required criteria of the qualifications and experience, (c) Final selection of a candidate through interview. It is also needed in the scenario of talent management vis-à-vis identification of the experts in various technical domains for interviewing the candidates, and recommendation of career development programmes (higher qualification and training needs) for employees.

A solution has been suggested (Figure 6) for implementation within organizations, which can be integrated with the HR Management processes to address the organization's strategic talent needs [5].

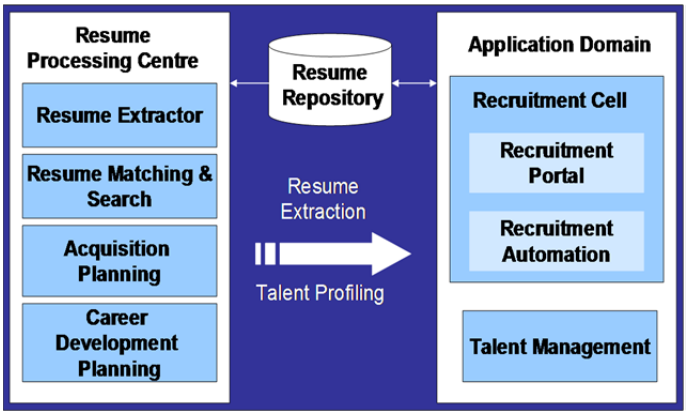


Figure 6. Framework for a centralized Resume Processing System

The resume extractor module extracts the resume data and the matching module searches for the right resume to match job requirements. These two modules can make use of a text mining framework like DISCOTEX (**Disco**very from **T**ext **Ex**traction) [1] to unearth the trends/patterns and profiles of talent available (Talent Profiling) in various organizational and external entities. DISCOTEX makes use of templates with pre determined set of slots to be filled by strings taken from the documents (job advertisements or resumes). The fillers for the slots in a particular template are extracted from the document. Figure 7 shows a sample text advertisement for a post and its filled template where several slots may have multiple fillers.

Machine learning systems like RAPIER (Robust Automated Production of Information Extraction Rules) [6], BWI (Boosted Wrapper Induction) [7] or any other high accuracy learning system tailored to the specific needs of the domain can be used to automatically construct information extractors for a job advertisement as well as the resumes received in response to it. With the help of IE, a structured database of the extracted slots can be mined with standard data mining techniques to discover interesting relationships and patterns.

Sample text of Job Advertisement:

Post: Scientist 'B'

Location: VRDE, DRDO, Ahmednagar

The individual should possess Bachelors degree in Production/Mechanical/Automobile Engineering and at least 2 years Experience/Specialization in Production of wheeled and tracked vehicles/ systems. The upper age limit as on closing date of the Advertisement is 28 years.

Filled Template:

Post: Scientist 'B'

Location: VRDE, DRDO, Ahmednagar

Required Qualification: Bachelors degree in Engineering

Required Subject: Production/Mechanical/Automobile Engineering

Required Experience /Specialization: Production of wheeled and tracked vehicles/ systems

Required years of experience: 2+

Age limit: 28 years

Figure 7. Sample text of a job advertisement and filled template

Besides, the rules mined from this structured database can be used to predict additional information to be extracted from the future resumes. Each rule represents a chunk of knowledge describing the relationships between slot values as a conditional statement, commonly in the form of a left hand side (LHS) consisting of a conjunction of several conditions and a single right hand side (RHS) term. Figure 8 shows the sample rules portraying the propensity of occurrence of certain slots (RHS of rule) in relation to other slots (LHS of rule).

- 'DIPR' ∈ *Establishment* and 'Personnel Selection' ∈ *Specialization* ⇒ 'Psychology' ∈ *Discipline*
- 'Central Drug Research Institute' ∈ *Institute* ⇒ 'Pharmacology' ∈ *Discipline*
- 'Robotics' ∈ Specialization ⇒ 'CAIR' ∈ *Establishment*
- 'Nuclear Medicine' ∈ *Discipline* and 'Hyperthyroidism' ∈ *Specialization* ⇒ 'MBBS' ∈ *Degree*
- 'Automobiles' ∈ Specialization and 'Vehicles' ∈ *Specialization* and 'Production' ∈ *Specialization* ⇒ 'Mechanical' ∈ *Discipline*
- 'Scientist B' ∈ *Present-Rank* and 'Improving personal qualities' ∉ *Career-Development-Preference* ⇒ 'Improving subject knowledge' ∈ *Career-Development- Preference*

Figure 8. Production Rules mined from extracted database

Adopting this system will eliminate the cumbersome filling of application forms online and a candidate will only need to upload his resume on the recruitment portal. The HR Manager will also just need to define the requisite criteria instead of employing conventional filtering techniques to sieve the information from the uploaded resumes.

IV. LIMITATIONS

In order to develop the prospective resume processing system, it is essential to provide it a knowledge base. A comprehensive domain dictionary consisting of the vocabulary in respect of the relevant domains can improve the system's response significantly. Table 1 is an example demonstrating identification of words belonging to parent category 'Computer Science' and sub-category 'Artificial Intelligence'.

Parent Category	Sub- Category	Words
Computer Science	Database Management Systems	SQL, Query Optimization Oracle, Database Design …
	Artificial Intelligence	Natural Language Processing, Pattern Recognition, Robotics, Genetic Programming, Heuristics..
	Cyber Security	Cryptography, Encryption, Firewall..
	Graphics	Animation, Rendering, Imaging..

Table 1.Structure of the domain dictionary

Pulling together the right knowledge into a domain dictionary is an onerous job. Knowledge for such systems can be derived from expert sources like experts in the given field, journal articles, reports, databases and so on. The system can demonstrate some intelligence, depending on the content and quality of its database.

V. SUMMING UP

This paper provides an overview of some technologies used for mining of resume repositories such as information retrieval, natural language processing, text mining and data mining. The contemporary advances in these technologies are making the automated processing of resumes more accurate for making timely inferences and decisions to manage talent in an organization. The advantages of implementing the proposed centralized resume processing system within an organization are apparent in its ability to better analyze resumes for tapping the best talent available from the external talent pool. The acquisition of the requisite domain knowledge for the potential system remains as one of the bottlenecks in building it. Developing such a system for an organization will be a very big step forward.

REFERENCES

[1] Raymond J. Mooney and Un Yong Nahm, Text Mining with Information Extraction, Proceedings of the 4th International MIDP Colloquium, September 2003

[2]http://www.i5.informatik.rwthaachen.de/lehrstuhl/projects/DocMINER/DocMINER.html, 2004.

[3] R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. ACM Press, New York, 1999.

[4] Nicholas Kushmerick, Wrapper Induction, Efficiency and effectiveness, Elsevier,Artificial Intelligence 118(2000) 15–68.

[5] R. Srivastava, G. K. Palshikar, RINX: Information Extraction, Search and Insights from Resumes, Proc. TCS Technical Architects' Conf., (TACTiCS 2011), Thiruvanthapuram,India, Apr 2011.

[6] M. E. Califf and R.J Mooney. Relational learning of pattern-match rules for information extraction. In Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99), pages 328–334, Orlando, FL, July 1999.

[7] D. Freitag and N. Kushmerick. Boosted wrapper induction. In Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), pages 577–583, Austin, TX, July 2000. AAAI Press / The MIT Press.

MANOGNA .N is pursuing her final year in Masters Degree of Information Technology from Jawaharlal Nehru Technological University, Hyderabad. She is presently working as Junior Research Fellow (JRF) in the Recruitment and Assessment Centre, DRDO, Delhi. Her areas of interest include Information Retrieval, Artificial Intelligence, and Advanced Computing

SUMEDHA MEHTA has obtained her Masters degree in Computer Applications from Indira Gandhi National Open University. She is presently working as Technical Officer 'A' in the Recruitment and Assessment Centre, Delhi. Her areas of interest include Information Retrieval and Software Engineering.