International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Research Article

Relevance Ranking Algorithm for Job Portals

Vasudha Sarda[Å*], Prasham Sakaria[Å] and Sindhu Nair[Å]

[Å]Computer Science, D.J.Sanghvi College of Engineering, Mumbai, India

Accepted 07 Sept 2014, Available online 01 Oct 2014, **Vol.4, No.5 (Oct 2014)**

Abstract

Hiring an appropriate candidate for a job profile which is supposed to perform highly in a company is a matter of concern for companies' management. Job portals have a number of resumes for a particular job posting from which the company needs to shortlist the qualified applicants. In this paper, the applicants for a job profile will be ranked based on their resume and social media presences. The ranking process will be based on the attributes which influence the performance of the employee of the company. The attributes are evaluated using classification algorithm like Decision Tree and Naïve Bayes. There will be a match making system built where the companies will be given a list of ranked candidates using information retrieval technique like two way relevance ranking.

Keywords: *Naïve Bayes, Decision Tree, Two way Relevance, Performance Evaluation, Screening Candidates, Information Retrieval.*

1. Introduction

Companies get a number of resumes for a job posting which is manually analyzed and either accepted or rejected based on how appropriate a candidate is for a job opening. This tedious job can be replaced by a resume ranking process. An Information Retrieval technique can be applied to rank the resumes based on a number of factors. But each company has different attributes which contribute to its employees performance. Thus evaluation of the employee performance is carried out for a specific company using data mining techniques to find out the attributes. These attributes can be further used for matching of a candidate to a job profile posted by that specific company. Based on the matching process using two-way relevancy ranking the candidates which do not qualify are filtered out. The company is given a list of candidates which qualify for the position with a score found out by the ranking algorithm. Social media profiles are also viewed while ranking the candidates. Thus the external data from the candidate's profiles on sites like LinkedIn can be used to improve the efficiency of relevance ranking.

2. Literature Review

The innovation in the communication technologies has brought about a major change in all spheres of our life. The internet has changed the way people communicate and gather information. The changes made in traditional recruitment process by hiring using corporate career pages are advantageous. (De Meo *et al*., 2007). It helps in targeting a wider range of candidates for a position for a lesser amount. The time consuming process of manually scrutinizing all the resumes has been replaced by E-recruitment, an automated process of assessing the profiles of the applicants. This has reduced the cost per profile as they do not need a special agent to help them find a suitable candidate.

Human Resource Department is under immense pressure by the company's management to recruit the candidate most suitable for a company and thus a change in its functioning will contribute to an organization strategically. (Parry and Tyson, 2009).

There have been a number of job portals available today which have E-recruitment tools. The companies post job openings on these portals and people looking for jobs post their profiles. The matching quality still inadequate for achieving a good match between the job description and candidate profiles. Chein and Chen (2006) have tried to improve the candidate selection process by building a model using data mining techniques. Based on the attributes selected the performance of the candidate can be predicted looking at their resumes, job applications and interviews. This can help to take the decision of whether or not to employee the candidate.

In Malaysian higher institutions data mining techniques have been applied on their application as well as their past experience to predict his talent using the given information. Jantan *et al*. (2010) .

The E-recruitment sites suffer from matching the job profile to the resume of the candidates. (Bizer *et al*., 2005). Basic theory and mathematical tools are already available; however, the most complicated part in job matching process is the matching between the candidate's information and employers' requirement. Matching

*Corresponding author: **Vasudha Sarda**

algorithm has to be flexible to the priority to the employers and candidates but at the same time form a balance between employers and candidate's preferences. (Terzis and Economides, 2005). Job matching process involved the calculation of similarity between candidate profile and job requirement specified by the employers. The recent developments have led to increase in the use of job portals by both the recruiting heads as well as the employees. (Gueutal and Stone, 2005; Ross and Young, 2005).

Techniques such as Decision Tree, Rough Set Theory, and Naïve Bayes could be applied on various factors which were considered as attributes for a particular company. 'Improving Job Search by network of professionals and companies' a research paper where there a clusters formed between categories of professionals and the various requirements by the companies. It is used to discover important relationships between companies and professionals.

While the 'Semantic Web-based', used by Monster.com uses the data exchanged between employers, applicants and job portals describing the occupations and job skills. 'Agent-based Application for Supporting Job Matchmaking for Teleworkers' is a system which is used to find the time consuming process of finding a working partner in a telecommunication sector.

3. Implementation

Machine learning techniques are used to rank the resumes based on the job profiles which are posted on the websites. It is divided into two parts, first being the relevancy attributes which affect the selection process the most. These features influence the matching of the job profiles to the companies the most. These attributes may vary from company to company. Generally these attributes are the same as the ones which influence the performance of the employees for a company. Thus we need to first find those attributes by evaluating the performance of the employees of a company thus while recruiting they can predict the performance of a candidate. This is done using classification techniques like Decision Tree or Naïve Bayes algorithm which can test the attributes which contribute to the performance of an employee. This was implemented using CRISP which consists of five steps which include: Business understanding, data understanding, data preparation, modelling, evaluation and deployment.

A. Data Classification Preliminaries

Classification of the data is divided into two steps. The first step being, classification applied on the training data. Each instance is assumed to belong to a predefined class. In the second step, the model is tested using a different data, test data set that is used to estimate the classification accuracy of the model. At the end, the model acts as a classifier in the decision making process. The C4.5 technique, which is one of the most powerful Decision Tree technique, can produce both decision tree and rule-sets. C4.5 creates an initial tree using the divide-and-conquer algorithm.

WEKA toolkit (Witten *et al*., 2011) is a widely used toolkit for machine learning and data mining originally developed at the University of Waikato in New Zealand. It contains a large collection of state-of-the-art machine learning and data mining algorithms written in Java. WEKA contains tools for regression, classification, clustering, association rules, visualization, and data pre-processing. WEKA toolkit package has its own version known as J48. J48 is an optimized implementation of C4.5 rev. 8.

B. Data Collection Process and Data Understanding:

When the idea of the study came in to mind, it was intended to apply a classification model for predicting performance depending on a dataset from a certain IT company. So that any other factors regarding the working environment, conditions, management and colleagues would have similar effect on all employees. The data which is required for experiment was collected using questionnaires. The employee of a particular IT company had to fill the questionnaires which had attributes which were considered to influence the working pattern of an employee.

C. Data Preparation

The first step to prepare the data is to collect the questionnaires. The data in the questionnaire is transferred to an Excel Sheet. The attributes were modified in order to apply the classification algorithm. Some attributes like experience years and service period, have been entered in continuous values. So, they were modified to be illustrated by ranges. Other attributes like specialization, job title and rank, have been generalized to include fewer discrete values than they already have. These files are prepared and converted to (arff) format to be compatible with the WEKA data mining toolkit (Witten *et al*., 2011), which is used in building the model for the different classification algorithms.

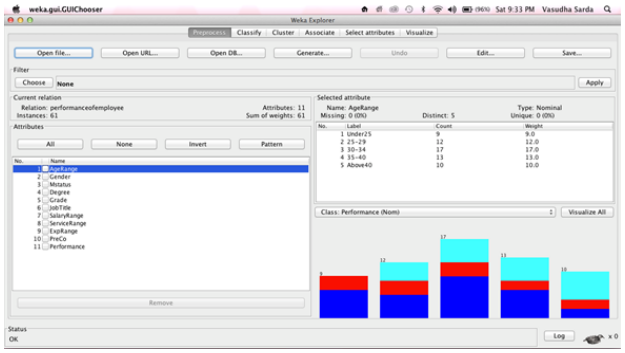

Fig 5.1 showing Pre-processing Stage in performance evaluation of employees

D. Evaluation and Deployment

The two data mining techniques taken to predict the performance of the employee are Decision Tree technique, C 4.5 and Naïve Bayes theorem. The experiments are

carried out using 10 fold cross validation technique which is used to estimate the accuracy of both the techniques. We have taken 13 attributes and 61 instances for both the algorithms. The attributes being Age, Gender, JobTitle, SalaryRange, Grade, Degree, ServiceRange, ExpRange, PreCo and Performance. The Experience Range has the maximum gain ratio, which made it the starting node and most effective attribute. Other attributes participated in the decision tree were SalRange, ExpYears, Grade, Age, Gender other attributes such as: Grade and Service Range were the other nodes of the decision tree.

The tree indicated that each attribute had an effect on the performance of the employee but the most effective ones were- Grade, Experience Range and Service Range. Other hints could be extracted from the tree indicates that young employees have better performance than older ones. Wherever Gender is taken into consideration, Males have higher performance than Females. Moreover, employees with higher graduation grades have higher performance. Finally, employees with higher ranks have less performance indicating that managers work less than less ranked employees.

The decision tree built using the C4.5 algorithm was so much pruned to consist of only three attributes; with ExpRange as the starting node, ServiceRange and Grade as other attributes. J4.8 forms a pruned tree having 8 leaves and the size of the three is 10. Experience sometimes affects the employee's motivation and therefore performance in a positive way. Such, personality factors can be recognized by decision makers in interviews, so that they can complete their knowledge about the applicant. The university general specialization has a very close effect to performance as the job title. This could be due to the relationship between these two factors.

The Naive Bayes algorithm had less number of correctly classified instances as compared to the J4.8 algorithm taking the same number of attributes and instances into consideration. It had 34 instances classified correct from the 61 instances, thus having a lower performance than J4.8.

When some of the attributes with less gain ratio like AgeRange, Gender and MStatus were removed from the pre-processing part and then the data was classified the efficiency of Naive Bayes algorithm was unchanged but the efficiency of J4.8 increased. Thus the number of attributes can have different effects on different algorithms. Some personal information like age, marital status and gender also affects the performance. Nevertheless, the age has not clear effect on the performance, since sometimes the performance increases with age, which adds the experience factor, other times, it decreases showing the highest motivation with the younger employees. Marital status, on the other hand, is clearer in its effect, since single employees in all experiments have shown better performance from married employees and even much better than married with kids employees.

The performance of an algorithm can be calculated using a visualization layout is done using confusion matrix or contingency table. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. Precision is the fraction of retrieved instances that are relevant, while recall which is also known as sensitivity is the fraction of relevant instances that have been retrieved in binary classification. Measure of relevance is estimated using precision and recall. Precision calculated using Decision tree classification algorithm is more than the precision calculated by Naïve Bayes and the recall of Naïve Bayes is more than the one calculated using Decision Tree algorithm.

Table 5.1 showing the comparison between the methodologies used for evaluating the performance of the employees

Decision Tree	Naive Bayes
Correctly classified instances: 63.93444%	Correctly classified instances: 55.7377 %
Precision is 0.667 having 333 relevant data missed.	Precision is 0.636 having 364 relevant data missed.
Recall is 0.929 having 71 noisy data.	Recall is 0.750 having 250 noisy data.
After removing attributes like Gender, Age and Mstatus the correctly classified instances increased to 68.8525 %	After removing attributes like Gender, Age and Mstatus the correctly classified instances remains the same.

4. Relevance System

A training data is used to build a model on which the ranking algorithm needs to be implemented. So resumes are graded by Human Resource executives based on a particular job description. It was first determined whether the resumes of the candidate met the minimum threshold and could be passed forward for an interview process. The same resumes were given to a number of evaluators to get a wholesome review and it was found that they had given the same rate to most of the resume.

Matching of the resumes to a job description was done based on two types of data, internal data and external data. Internal data were the previously found attributes which affected the performance of an employee like skills for the simple matching. There can be other sophisticated features can also be used like employment and job-hopping, if an applicant is overqualified, previous versus current salary expectations, career trajectory, company prestige, if an applicant previously worked for a competitor, required and desired skills, certifications, school rank, education timeline, several different semantic relationships between the resume and job description, resume and job description spectral density, social imprint, company connections, social network size, personality traits, cognitive profile, unique analysis of data from the Bureau of Labor and Statistics and many other available sources, SEO.

We use two-way relevancy system where we first find all the requirements of a company which it needs in its employee for a particular position then it finds all the requirements a candidate wants from a company. The given set of candidates which pass the filtering process mentioned above seem to be a better match to the job requirement. The requirements placed by the company is

matched with the requirements posted by the employee, like if the company is a Startup or a MNC, the pay given by the company or even the location of the company can be some of the requirements of the candidate which if the company's profile doesn't fulfill it will not be passed on to the candidate.

Matching of the candidates and companies can be done using Gradient Boost Decision Trees machine learning algorithm . It is an algorithm which is commonly used for ranking. Yahoo uses variants of this algorithm for its machine learning ranking engines. It is generally used for classification problems by reducing them to regression, by producing a weak model like decision trees.

External data taken from social media websites as well as public database could also improve the quality of the matching process. LinkedIn, a professional websites can give more information about the candidate apart from the data available on their resumes. The Recommendation System is one of the key features of a profile on LinkedIn.

The behaviour of the candidate on the social media site can have an impact on their performance. The data available on social media profiles can be add to the limited amount of data available on the resume to help distinguish between candidates suitable for a job and not suitable for it. The information available on these sites can be utilized for the machine learning job-candidate algorithm where say the number of recommendations received by a candidate can be taken as one of the attributes used for the ranking of the resumes by assigning it some weight. The candidates which are applying for some job profile is made to register on the job portal by signing in with LinkedIn so that we can access the job profiles of the candidate as well as its connections used for the matching process.

Conclusions

We can conclude by saying that attributes like age and gender have a mixed effect on the performance of the employee and we cannot take these attributes alone cannot be taken into consideration for predicting the performance. Experience Range, the number of years or experience gained by the employee along with Grade received in college and the Service Range, the number of years spent in that field have considerable effect on the experiment. The resumes can be ranked in companies of the candidates during the selection section based on these attributes so they can accept only those employees who will be able to perform well in that environment.

The two-way relevance ranking process is time consuming and is not proven to be efficient for all matching systems. Thus query independent ranking has been found to be the simplest and easiest way of ranking the resumes. There needs to be an improvement in the selection of the attributes for matching and those attributes should be considered which have found to be to increase relevance accuracy.

References

Qasem A. Al- Radaideh Eman Al Nagi, (2012) Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance, International Journal of Advanced Computer Science and Applications, Vol. 3, No. 2.

Sexton, R.S., McMurtrey, S., Michalopoulos, J.O., Smith, A.M. (2005). Employee Turnover: A Neural Network Solution, Computer & Operations Research, 32, pp. 2635-2651.

Witten I. Frank E., and Hall M. (2011). Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition, Morgan Kaufmann Publishers.

Fernando Diaz, Donald Metzle (July 2010), Relevance Ranking in Online Dating Systems, SIGIR'10, Geneva, Switzerland.

Salathiel Bogle, Suresh Sankaranarayanan(May 2012). The Job Search System In Android Environment-Application Of Intelligent Agents, International Journal of Information Sciences and Techniques (IJIST) Vol.2, No.3.

Jacob Bollinger, David Hardtke, Ben Martin (October2012), Using Social Data for resume Job Matching, DUBMMSM'12, Maui, Hawaii, USA.

Amit Shah, Rose Catherine, Karthik Visweswariah, Prospect: A System for screening Candidates for recruitment, CIKM'10, October-26-30, 2010, Toronto, Ontario, Canada.

Ralf Mikut, and Markus Reischl, (2011) Data mining tools Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Vol. 1, Issue 5.