iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** III **Month of publication:** March 2020

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue III Mar 2020- Available at www.ijraset.com

Job Recommendation System using Resume Data Extraction

Prof. A. A. Bamnikar[1], Ranjit S. Jeve[2], Divya S. Nair[3], Siddhi A. Nalage[4], Rachana D. Chavan[5]
[1]*Assistant Professor,* [2, 3, 4, 5]*Student, Department of Computer Engineering, PDEA COEM, India*

Abstract: ***There is a tremendous increase in online job recruitment, traditional methods of hiring candidates for job have become of no use. The knowledge extracted from applicants resumes or documents is required for mainly two reasons one to support the automated selection of candidates, and second to efficiently route them to their corresponding different categories of job within which they're suitable. This ends up in minimizing the hassle required by HR and employers to manage and organize resumes, also to screen the candidates who are insignificant for the post. We've to seek out the simplest way to avoid wasting the time of the candidates which they earlier spent on searching for suitable job and reducing the complexity of screening process. This can increase the productivity and efficiency of the general recruitment process. The recommended results can do higher precision, and that they become more relevant with users' choice. In this paper we've addressed need of recruiters and candidates, where this project helps the final year students to seek out the right job of their required skill-set which also matches with the corporate profile. Students just have to upload their resumes in pdf, image or any other form. Our interface will extract the information from the resumes using Optical Character Recognition.***
Keywords: OCR, Threshold, Gaussian blur, Tokenization.

I. INTRODUCTION

In India itself there are about 13 lakh university students graduating per year, coping with the large amount of recruiting information on the web, employment seeker always spends hours to search out useful ones. Finding and hiring the correct talent from a good range of candidates remains one in all the foremost important and challenging tasks of the HR department in any organization [3]. To handle this challenge, many companies have shifted to e-recruiting platforms [5,6]. These platforms reduce the value, time and energy required for manually processing and checking applicant resumes. As stated in [7], there have been over 32,000 e-recruitment sites in 2012 for helping jobseekers and recruiters worldwide. Consistent with the International Association of Employment websites (IAEWS) [8], the quantity of e-recruitment systems has become over 60,000 in 2019. To scale back this tedious work, we design and implement a recommendation system.

The recommender systems will help to determine the interested items for a selected user by employing a range of data resources that's associated with users and items. It provides a hybrid approach to classify resumes and their corresponding job post by utilizing an integrated occupational categories of information base. The exploited knowledge domain assists in classifying resumes and job offers under their corresponding occupational categories. Many researches in industry and academics are known to develop new approaches for recommender systems within the last decade. Such approaches try to match terms in CV descriptions to job position descriptions. During this work a unique approach is tailored within the sense that the semantic matching primarily concerns the applicant skills as denoted within the respective LinkedIn profile descriptions [4]. Recommender systems are being broadly accepted in various applications to suggest products, services, and knowledge items to customers. Many e-commerce applications join recommender systems so as to expand customer services, increase selling rates and reduce customers search time. For instance, a good range of companies like the web book retailer Amazon.com, books, and news articles. Additionally, Microsoft provides users many recommendations like the free download products, bug fixes etc. These companies have successfully founded recommendation systems and have increased web sales and improved customer fidelity. Moreover, many software developers provide stand-alone generic recommendation technologies. The highest providers include Net Perceptions, Epiphany, Art Technology Group, Broad Vision, and Blue Martini Software. The previous couple of decades have witnessed a very impressive growth of data across the web. The massive information is unused across the globe and it requires rigid methodology to mine and extract the text. The expansion of data is increasing rapidly, and it becomes more important to detect useful pattern from the info [2].

II. RELATED WORK

Many organizations today are pushed to implement flexible organizational and dealing structures like team- or project-based working modes the necessity to develop such decision support among others arises from the very fact that information technology in the past decade has changed the ways people collaborate. Many approaches and techniques have been proposed for addressing the e-recruitment process. In this context, some approaches attempt to overcome issues associated with the matching process between the persons resumes and their corresponding job offers, while others attempt to classify resumes and job posts before starting the matching process [16,13]. For instance, the authors of [15] have proposed an approach for the automatic matching and querying of information in the human resources domain. The proposed approach exploits DISCO, ISCO and ISCED taxonomies to achieve better matching results than traditional techniques that simply overlap keywords between the content of job posts and the candidate's resume ignoring the hidden semantic dimensions in the text of both documents [2].The proposed system automatically generates classification rules from a set of pre-classified job openings and assigns one or more class for each job post. The main drawback of this system is that some taxonomies doesn't cover the occupational information that is more relevant to the modern workplace [10]. Other systems utilize machine learning algorithms in order to comment segments of resumes with the appropriate category, taking the advantage of the resume's contextual structure where related information units usually occur in the same textual segments [13, 16]. However, the main drawback of these approaches is that a large part of the produced results suffer from low precision since the information extraction process passes through two not so strong stages, in addition to the time needed to pre-process and post-process job posts in order to reduce the error and increase the classification accuracy

III. METHODOLOGY

Image Pre-Processing

Following are the steps of image pre-processing:

1) Loading a Resume Image.
2) Converting Image from BGR to GRAY.
3) Applying threshold to the image.
4) Applying Filter to the thresholder image.

A. Loading an Image

In this step an image is loaded by the program. When the image gets loaded it is in the form of matrix. This matrix is stored in the variable. Now this variable will act as a image which is further used to carry out the required operations. The values inside this matrix are the pixel intensity at a particular point. The particular pixel intensity has 3 channels namely - BLUE, GREEN, RED. This is because every colour in the world can be represented by combination of these 3 colours.

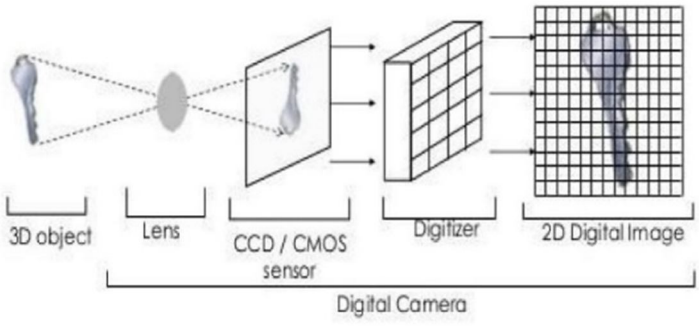


Figure 1: Formation of Digital Image

This diagram clearly shows the pipeline of formation of an image and how the image is stored. Thus, all the operations and manipulations that take place on an image are done pixel by pixel. Where each pixel value is taken and passed through the different mathematical equation to get the desired output

B. Converting Image BGR to GRAY

If each colour pixel is described by a triple (R, G, B) of intensities for red, green, and blue, and uses different algorithms to convert to grey. On converting the image to grey the image which was earlier 3 channels is converted to 1 channel. The main reason behind converting an image to grayscale is that all the thresholding, filtering, edge detection and pre-processing algorithms work only on single channel images. The GIMP image has following 3 algorithms for converting and colour image to grayscale. The lightness method averages the most prominent and least prominent colours: (max (R, G, B) + min (R, G, B)) / 2. The average method simply averages the values: (R + G + B) /3. The luminosity method is a more practical version of the average method. It also averages the values, but it forms a weighted average to account for human approach. We're more sensitive to green than other colours, so green is weighted most heavily. The formula for luminosity is 0.21 R + 0.72 G + 0.07 B. Original Lightness Average Luminosity

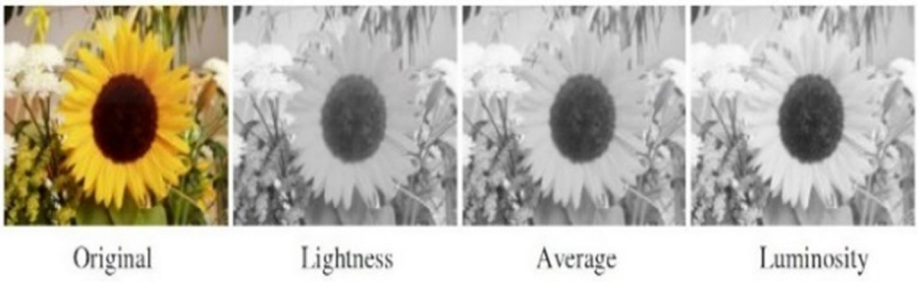


Figure 2: Comparison of different grayscale methods

The lightness method tends to reduce contrast. The luminosity method works best overall and is the default method used. However, some images look better using one of the other algorithms. And sometimes the three methods produce very similar results.

C. Applying Threshold to Image

Thresholding is simplest method used for image segmentation. From a grayscale image, thresholding can be used to create binary images. The simplest thresholding methods which replace each pixel in an image with a black colour pixel if the image intensity I(i,j) is less than some fixed constant T (which is. I(i,j)<T), or a white pixel if the image intensity is greater than that constant. In the example image on the right, this results in the dark tree becoming completely black, and the white snow becoming completely white. The input to thresholding operation is normally a grayscale or colour image. In the simplest implementation, the output is a binary image which represents the segmentation. Black pixels in image correspond to background and white pixels correspond to foreground (or vice versa). In simple implementations, the segmentation is determined by a single parameter the intensity threshold. In each single pass, each pixel in image is compared with this set threshold. If the pixel's intensity is more than the set threshold, the pixel is set to, say, white in the output. The output is displayed as per the threshold limit. If it is less than the threshold, it is set to black. In more sophisticated implementations, multiple thresholds can be specified, so that a band of intensity values can be set to white while everything else is set to black.

There is algorithm which calculates the threshold for small region of image and that process of calculating is known as Adaptive Thresholding.

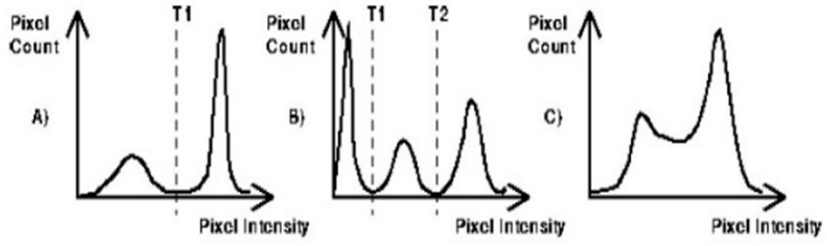


Figure 3: Thresholding using histogram

Due to Adaptive Thresholding, we get different thresholds for different regions of the same image and it gives us better results for images with different illumination.

D. Filtering Image

In image processing, a Gaussian blur is the result of blurring an image by a Gaussian function. Gaussian function is widely used in graphics software, typically it is used to reduce image noise and reduce the other details. The visual effect of this blurring technique is a smooth blur resembling which is viewing the image through a translucent screen, distinctly different from the bokeh effect produced by an out-of-focus lens or the shadow of an object under usual illumination. In two 2D, it is the product of two such Gaussian functions, one in each dimension:

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}} \qquad (1)$$

E. Text Extraction from Image

Optical character recognition (also optical character reader, OCR) is the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, or a photo of document.

OCR is widely used as a form of information entry from printed paper data records – whether it is passport documents, invoices, bank statements, computerised receipts, mails, or any suitable documentation. OCR is a common method of extracting printed texts so that they can be edited electronically, searched, stored more compactly, and used in machine processes such as cognitive computing, machine translation, text-to-speech, key data and text mining. OCR is a field of research in pattern recognition, artificial intelligence and computer vision techniques.

1) *Loading The Image File:* The very first thing to support OCR process is it must support a wide range of file formats, including PDF, BMP, TIFF, JPEG, and PNG files. Once the file is uploaded, the software can begin to work. These files can be scanned documents, photographs, or even read-only files. Once the files are loaded OCR software will transform these files into editable (word format) data.
2) *Improving Image Quality and Orientation:* Depending on the method in which the image file was created, there are a number of issues that may arises such as noise and sometimes uncertainty in data. More often than not, an image file will be skewed or contain "noise". In this stage of OCR, the software will remove any "noise", and improve the overall quality of the images. This is a critical step work to de-skew, remove noise, and improve the overall quality of the images. This is a critical step as blur or skewed images are not interpreted properly.
3) *Removing Lines:* Lines can prove to be dangerous when interpreting characters. To maintain the accuracy of data possible lines are detected and removed. This allows for better recognition quality when converting tables, underlined words, etc. It makes sure that data is not lost. The importance of image quality, the removal of lines will ensure that characters are recognized accurately.

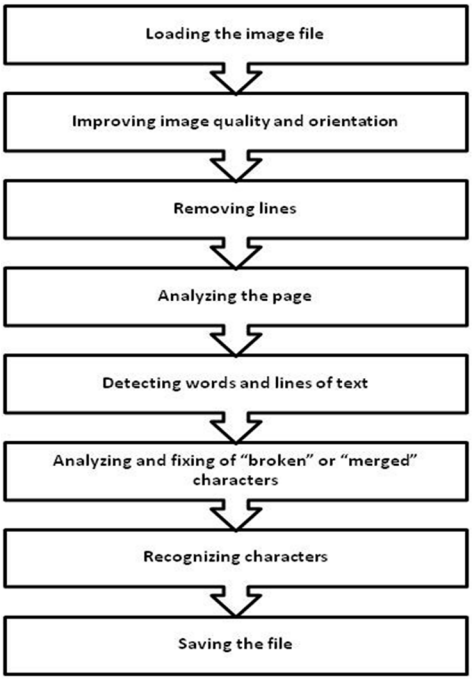


Figure 4: Flowchart of OCR

4) *Analyse the Page:* During this stage of Optical Character Recognition, the layout of the original file is noted. This includes the detection of text positions, white spaces, and the prioritization of significant text areas or sections.
5) *Detect Words and Lines of Text:* This is the stage where character recognition is started. The software begins to identify each words and entire lines of data. This is a critical pre-process for properly recognizing characters as it sets the stage to analyse and correct broken or merged characters.
6) *Analyse and fix the "Broken" and "Merged" Character:* Depending on the quality of the original file, there are often errors where characters are broken or blurred. The OCR software has to break down and resolve these errors in order to properly interpret the appropriate characters.
7) *Recognizing Characters:* This is the most important function of OCR. Now that the original file has been cleaned, processed, and fixed – the OCR technology can start to read and translate these characters. Each image of every character is then converted into a character code which are unique. If the algorithm is not sure of a character the software will produce multiple character codes and choose the proper character later on.
8) *Saving the File:* After the file is fully made clear, it can be saved to your desired format. While there is much to this OCR software, these 8 steps are the primary processes involved in OCR.

F. Feature Detection

Also known as intelligent Sharacter recognition (ICR) or feature extraction, this is a much more advanced way of spotting characters. Apply that rule and you'll recognize almost all capital letter As, no matter what font they are written in. Instead of recognizing the complete pattern of a letter let's say A, you're detecting the individual component features (such as angled lines, crossed lines, or whatever) from which the character is made. Most modern omni font OCR programs (ones which can recognize printed text in any font) work by feature detection rather than pattern recognition. Some use neural networks (computer programs that automatically extract patterns and think like human brain)

G. Making Sense of words

To make sense of the words which we have used we have used natural language processing library namely called as nltk, which is compatible with python 3.0 version. The library has built in functions which helps us to identify the words and get the meaning out of it. Thus, for the making sense we separate different words in the groups. This method of forming groups of words having a particular meaning is called the concept of tokenization. Tokenization is process of breaking the given text into small units which are called tokens. It can be words, or it can be numbers or punctuation mark. Tokenization method does this task by locating word boundaries. Last letter of a word and beginning of the next word is called word boundaries. Tokenization is also known as word segmentation.

There are mainly two types of tokenization:

1) word tokenization - This separate the sentence word by word. Thus it creates a list where a word from a sentence becomes an element of the list.

2) sentence tokenization - This separate the sentences sentence by sentence. Thus, it creates a list where a sentence from a group of sentences becomes an element of the list.

Figure 5: Structure of tokenization

H. Collecting Company Database

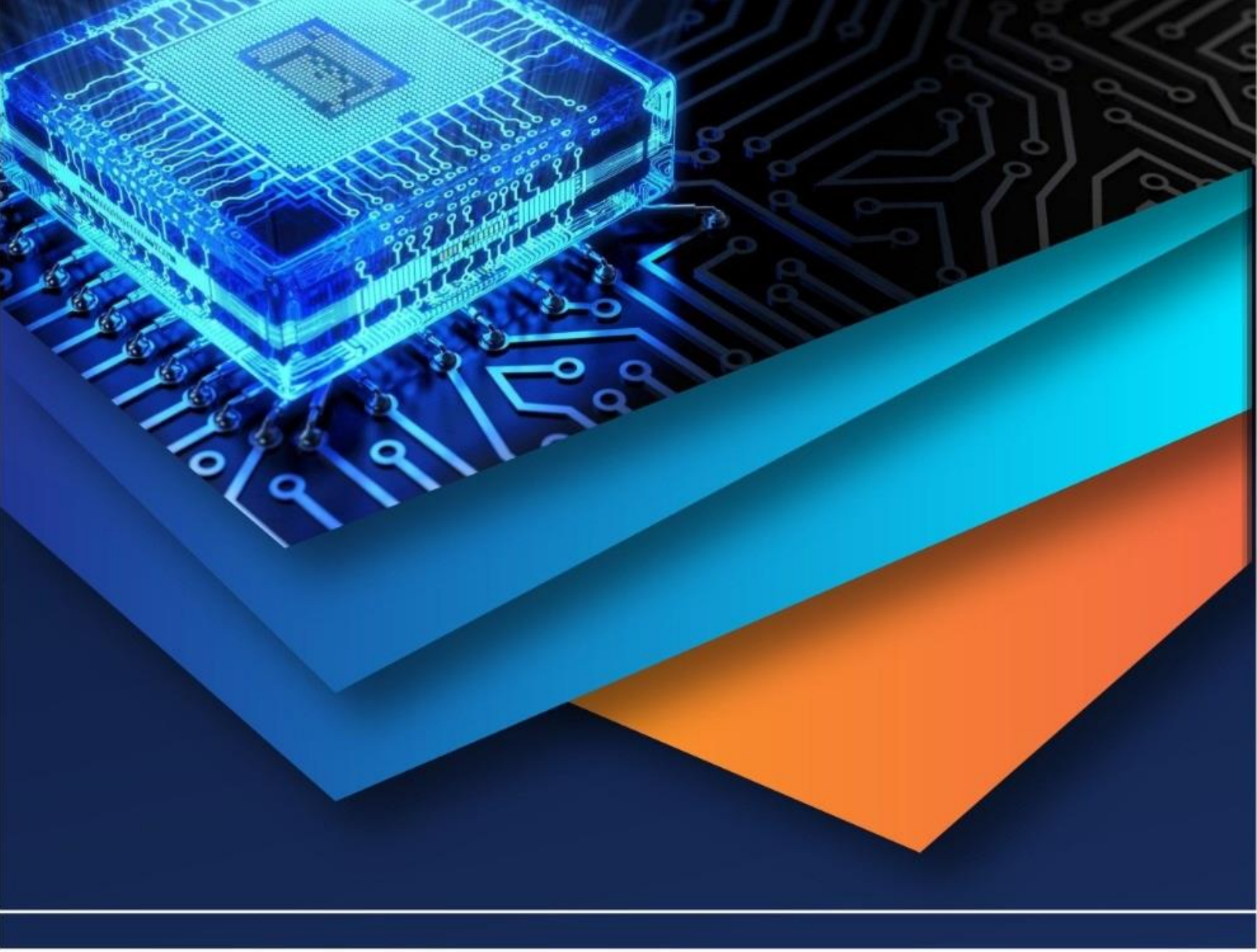
This is one of the challenging tasks of the project where we compiled the data of different companies on the basis of their job requirement. The data comprises of the company name, job description and the skill required to that specific job. This collection of data from the companies was survey conducted as part of the project. Where the different companies were searched online for their requirements, and the data was recorded. The skills required list of different companies was represented in the tabular form and afterwards it was passed to MySQL for the purpose of database handling and database management.

IV. CONCLUSION

The image pre-processing works with image format of the resume and OCR gets the texts from such images and PDF's. Using this systems number of things can be done. The speed of this operations is more than the manual work required. This kind of data then can be given to the filtering system. Which result to spending less time on filtering the job post's by candidates.

REFERENCES

[1] Abeer Zaroor, Mohammed Maree, Muath Sabha," JRC: A Job Post and Resume Classification System for Online Recruitment "2017 International Conference on Tools with Artificial Intelligence
[2] 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.
[3] J Chen, Z Niu, H Fu, "A Novel Knowledge Extraction Framework for Resumes Based on Text Classifier," Proceedings of the International Conference on WebAge Information Management. Springer International Publishing, pp. 540-543, 2015
[4] E Faliagka, L Iliadis, I Karydis, M Rigou, S Sioutas, A Tsakalidis, and G Tzimas, " On-line consistent ranking on e-recruitment: seeking the truth behind a well-formed CV," The Artificial Intelligence Review, 42(3), 515, 2014.
[5] T Schmitt, P Caillou, M Sebag, "Matching Jobs and Resumes: a Deep Collaborative Filtering Task," Proc. of the 2nd Global Conf. on Artificial Intelligence, pp.1-14, 2016.
[6] S Mehta, R Pimplikar, A Singh, LR Varshney and K. Visweswariah, "Efficient multifaceted screening of job applicants," Proceedings of the 16th International Conference on Extending Database Technology. ACM, pp. 661-671, 2013.
[7] S Al-Otaibi and M Ykhlef, "Job Recommendation Systems for Enhancing E-recruitment Process", in Proceedings of the International Conference on Information and Knowledge Engineering (IKE), Las Vegas Nevada, USA, pp. 433-439, 2012.
[8] The International Association of Employment Web Sites (IAEWS), avaliable from: http://www.icmaonline.org/international-association-ofemployment-web-sites, Date Visited: June 20, 2017
[9] Jayaraj, V., and V. Mahalakshmi. "Augmenting Efficiency of Recruitment Process using IRCF text mining Algorithm." Indian Journal of Science and Technology 8.16 (2015).
[10] 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.
[11] K Yu, G Guan, and M Zhou, "Resume information extraction with cascaded hybrid model." Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, pp. 499-506, 2005.
[12] F Javed, Q Luo, M McNair, F Jacob, M. Zhao, and TS. Kang, "Carotene: A Job Title Classification System for the Online Recruitment Domain," Proceedings of the IEEE First International Conference on Big Data Computing Service and Applications (Big Data Service), pp. 286- 293, 2015.
[13] R Kessler, N Bechet, M Roche, J. M Torres-Moreno, and M El-Beze, "A hybrid approach to managing job offers and candidates," Information Processing & Management, 48(6), 1124-1135, 2012.
[14] J.Martinez-Gil, A.L. Paoletti, and K.D. Schewe, "A smart approach for matching, learning and querying information from the human resources domain," In East European Conference on Advances in Databases and Information Systems, Springer International Publishing, pp. 157-167, 2016.
[15] M Fazel-Zarandi and M S Fox, "Semantic matchmaking for job recruitment an ontolgy based hybrid approach," In Proceedings of the 3rd International Workshop on Service Matchmaking and Resource Retrieval in the Semantic Web at the 8th International Semantic Web Conference, Washington D. C., USA, 2010.
[16] S Clyde, J Zhang, and CC Yao, "An object-oriented implementation of an adaptive classification of job openings," Proceedings of the 11th Conference on Artificial Intelligence for Applications, IEEE, pp. 9-16, 1995.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY