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
[1,2,3,4,5]Department of Computer Science
[1,2,3,4,5]PDEA's COEM, India

Abstract— Due to the increasing growth in online recruitment, traditional hiring methods are becoming inefficient. The extraction of structured information from applicant resumes is needed not only to support the automatic screening of candidates, but also to efficiently route them to their corresponding occupational categories. This assists in minimizing the effort required by employers to manage and organize resumes, as well as to screen out irrelevant candidates. We need to find a way to save the time of the students which they earlier spent on searching the companies and reducing the complexity of screening process. This will increase the productivity and efficiency of the overall process. The recommended results can achieve higher score of precision and recall, and they are more relevant with users' preferences before.

Keywords: OCR, Tokenization, Pattern Discover, Automatic Matching, Text Matching, Information Retrieval

I. INTRODUCTION

Considering only in India there are about 13 lakh university students graduating every year, dealing with the enormous amount of recruiting information on the Internet, a job seeker always spends hours to find useful ones. Finding and hiring the right talent from a wide and heterogeneous range of candidates remains one of the most important and challenging tasks of the HR department in any organization [3].

To address this challenge, many companies have shifted to exploiting e-recruiting platforms [5,6]. These platforms reduce the cost, time and effort required for manually processing and screening applicant resumes. As stated in [7], there were more than 40,000 e-recruitment sites in 2012 for helping jobseekers and recruiters worldwide. According to the International Association of Employment Web Sites (IAEWS) [8], the number of e-recruitment systems has become more than 70,000 in 2018. To reduce this laborious work, we design and implement a recommendation system. The latest technology designed to fight information overload is the recommender systems that originated from cognitive science, approximation theory, information retrieval, forecasting theories and related to management science and to consumer choice modeling in marketing. The recommender systems used to determine the interested items for a specific user by employing a variety of information resources that is related to users and items. We present a hybrid approach to classify resumes and their corresponding job post by utilizing an integrated occupational categories knowledge base. The exploited knowledge base assists in classifying resumes and job offers under their corresponding occupational categories. Many researches in industry and academic areas have been known to develop new approaches for recommender systems in the last decade. The interest in this area still remains high because it is composed of a problem-rich research area and has a wealth of practical applications.

Such approaches attempt to match terms found in CV descriptions to job position descriptions. In this work a different approach is adapted in the sense that the semantic matching primarily concerns applicant. skills as denoted in the respective LinkedIn profile descriptions [4]. Recommender systems are being broadly accepted in various applications to suggest products, services, and information items to latent customers. Many e-commerce applications join recommender systems in order to expand customer services, increase selling rates and decrease customers search time. For example, a wide range of companies such as the online book retailer Amazon.com, books, and news articles. Additionally, Microsoft provides users many recommendations such as the free download products, bug fixes and so forth. All these companies have successfully set up commercial recommender systems and have increased web sales and improved customer fidelity. Moreover, many software developers provide stand-alone generic recommendation technologies. The top providers include Net Perceptions, Epiphany, Art Technology Group, Broad Vision, and Blue Martini Software. The last few decades have witnessed a stupendous growth of information across the internet. The giant of information is 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 [2].

II. RELATED WORK

Many organizations today are pushed to implement flexible organizational and working structures such as team or project-based working modes the need to develop such decision support among others arises from the 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 challenges. In this context, some approaches attempt to overcome issues associated with the matching process between candidate resumes and their corresponding job offers, while others attempt to classify resumes and job posts prior to starting the matching process [11, 13, 15, 16, 18]. For instance, the authors of [16] 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 look for overlapping keywords between the content of job posts and the applicant'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 DOT doesn't cover the occupational information that is more relevant to the modern

workplace [17]. Other systems utilize machine learning algorithms in order to annotate 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 [11, 18]. However, the main drawback of these approaches is that a large fraction of the produced results suffer from low precision since the information extraction process passes through two loosely-coupled stages, in addition to the time needed to pre-process and post-process job posts in order to minimize the error and maximize the classification accuracy.

III. METHODOLOGY

A. Image Pre-Processing

Following are the steps of image pre-processing:

- Loading a Resume Image.
- Converting Image from BGR to GRAY.
- Applying threshold to the image.
- Applying Filter to the thresholder image.

1) 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.

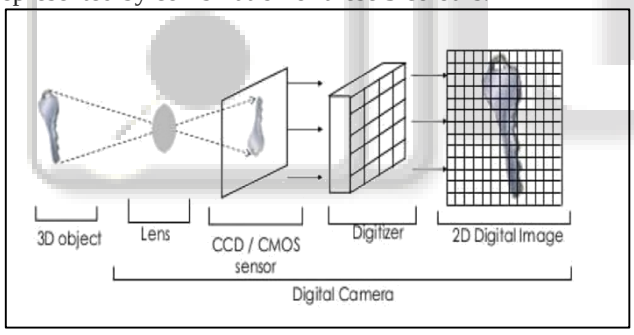

Fig. 1: Formation of Digital Image

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.

2) 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 sophisticated version of the average method. It also averages the values, but it forms a weighted average to account for human perception. 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

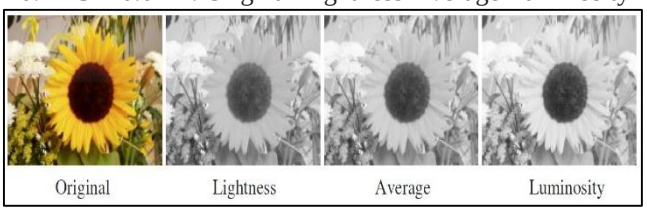

Fig. 2: Comparison of different grayscale methods

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.

3) Applying Threshold to image:

Thresholding is the simplest method of image segmentation. From a grayscale image, thresholding can be used to create binary images. The simplest thresholding methods replace each pixel in an image with a black pixel if the image intensity I(i,j)is less than some fixed constant T (that 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 a thresholding operation is typically a grayscale or colour image. In the simplest implementation, the output is a binary image representing the segmentation. Black pixels correspond to background and white pixels correspond to foreground (or vice versa). In simple implementations, the segmentation is determined by a single parameter known as the intensity threshold. In a single pass, each pixel in the image is compared with this threshold. If the pixel's intensity is higher than the threshold, the pixel is set to, say, white in the output. 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.

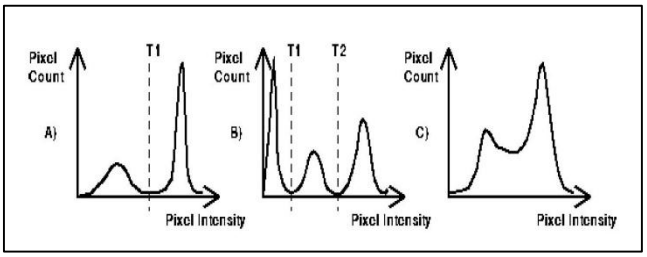

Fig. 3: Thresholding using histogram

There is method called Adaptive Thresholding which, in this, the algorithm calculates the threshold for a small region of the image. So, we get different thresholds for different regions of the same image and it gives us better results for images with varying illumination.

4) Filtering Image:

In image processing, a Gaussian blur is the result of blurring an image by a Gaussian function It is a widely used effect in graphics software, typically to reduce image noise and reduce detail. The visual effect of this blurring technique is a smooth

blur resembling that of 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 dimensions, it is the product of two such Gaussian functions, one in each dimension:

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

where x is the distance from the origin in the horizontal axis, y is the distance from the origin in the vertical axis, and σ is the standard deviation of the Gaussian distribution.

IV. CONCLUSION

The image pre-processing works with the image format of the resumes and the OCR gets the texts from such images and pdfs. There are lot of things which can be done using this system. The speed of such operations is more than the manual work. Such data then can be given to the filtering system. Which results into less time spent on filtering the job posts by candidate.

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 Web-Age 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-of-employment-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 (BigDataService), 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.
[17] About Occupational Information Network (O*NET). Available from: https://onet.rti.org/about.cfm. Date Visited: February 5, 2016.
[18] R Kessler, J Torres-Moreno, and M El-Beze, "E-Gen: automatic job offer processing system for human resources," in Proceedings of the artificial intelligence 6th Mexican international conference on Advances in artificial intelligence, Springer-Verlag: Aguascalientes, Mexico, pp. 985-995, 2007.

