IT Job Vacancy Recommender System using Random Forest Classifier

by Dewi Yanti Liliana

Submission date: 25-Jan-2022 08:48AM (UTC+0700)
Submission ID: 1747509731
File name: ob-vacancy-recommender-system-using-random-forest-classifier.pdf (407.07K)
Word count: 4669
Character count: 25396

Journal of Engineering and Social Sciences, Vol.1 No.1 (2021) p.23-30

IT Job Vacancy Recommender System using Random Forest Classifier

Dewi Yanti Liliana[1*], Vidi Ayuningtyas[1]

[1]Department of Computer and Informatics Engineering
Politeknik Negeri Jakarta, Jl. Prof. DR. G.A. Siwabessy, Kukusan, Kecamatan Beji, Kota Depok, Jawa Barat, 16424, INDONESIA

*Corresponding Author

Published online : 25 September 2021

To cite this article : *Dewi Yanti Liliana et al., Journal of Engineering and Social Sciences, Vol.1 No.1 (2021) p.23-30*

Abstract: The labor force participation rate in the last three years has increased in Indonesia. As many as 58% of workers looked for jobs online because the searching process was easier and faster than applying for jobs conventionally. Job seekers took a long time to find relevant jobs, ranging from 6 – 12 months. On the other hand, company recruiters were also looking for workers with experience relevant to their job vacancies. Usually, job recruiters check the candidate's experience through online media such as LinkedIn. The highest job vacancies are in the information and communication sector despite shrinking due to the pandemic in 2020. Hence, an automatic job vacancy classification system was needed to bring together recruiters and job seekers who met the requirements. To overcome this problem, this study aims to implement the Random Forest Classifier algorithm to classify jobs, especially in the information and communication sector. Jobs were classified into five categories: Data Scientist, Data Analyst, DevOps, Software Developer, and Mobile Developer. Data preprocessing step transformed data which was scrapped from LinkedIn website using API into features. Feature extraction was done using Word Embedding method for text analysis and Cosine Similarity to measure the level of texts similarity. The implementation used the Python programming language with the Flask framework to create a web-based application of job vacancy recommender system. Based on the experiment results, the accuracy was 94%. This application has been able to perform the job classification process and recommendations automatically with a promising result. Furthermore, job vacancy recommender system is applicable for the company to help them in a decision making with regard to an employment process.

Keywords: Job Vacancy, Machine Learning, Random Forest Classifier, Recommender System, Text Mining

1. INTRODUCTION

Employment is one of the fastest growing sectors worldwide. According to the Central Statistics Agency (BPS) (2018), the number of the workforce in Indonesia increased by 2.39 million people. In addition, the labor force participation rate increased by 0.18% so that the job search rate also increased. Job seekers usually look for jobs online because it is much easier than applying for a conventional job. According to Turczynski (2020), 58% of job seekers will look at job search websites. The search time is much shorter when compared to conventional searches. According to Business Insider (Lim, 2016), the average job seeker finds their new job within at least 6 months, whereas others only find a job within 12 months. This time is long enough to find a new job and can increase feelings of stress due to increasing financial pressure. In terms of recruitment, this time is related to the search for applicants who have relevant experience. Based on data released by the National Association of Colleges and Employers in 2019, 91% of recruiters are looking for workers who have experience and 65% of them are looking for workers with relevant experience. Usually, they check the

*Corresponding author: dewiyanti.liliana@tik.pnj.ac.id

candidate's experience through online media. According to Bitte's Recruiter Nation Report 87% of recruiters use LinkedIn to check on their job candidates (Bitte, 2016).

The information and communication sector itself will be dominated by a number of jobs. Job vacancies in the information and communication sector, especially Data Analyst, Data Scientist, DevOps, and Developer have a lot of interest. To find jobs that are relevant for applicants, it is necessary to create a job vacancy recommendation system in the information and communication sector so that job seekers get jobs in a shorter period of time. The recommendation system works by taking LinkedIn profiles to categorize jobs using the Random Forest Classifier algorithm. Random Forest Classifier is an ensemble learning method that performs classification by forming many decision trees at the training stage. Each tree is considered a classification and the classification output with weights is used for classification by majority voting (Wang, 2019). The research entitled "Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?" (Fernández-Delgado, 2014) shows that Random Forests consistently provide prediction accuracy among the highest compared to other algorithms on classification problems.

The initial step is the text analysis process which is carried out to measure the similarity between the profile and job description of job vacancies on other websites. This is done by using word embedding as in the study entitled "Sentences Similarity Analysis Based on Word Embedding and Syntax Analysis" (Xu, 2018). This study applies word embedding to find the level of similarity in a text. Word embedding works by converting sentences into vector form and using cosine similarity to measure the level of similarity in a sentence. This method is also known as doc2vec or document to vector. The research states that cosine similarity can be used to measure the level of similarity of the text resulting from word embedding. Values are sorted from highest to recommended to users.

This study aims to provide recommendations for suitable job vacancies based on experience and educational background. As well as helping to make decisions on the selection of the profession in the next career level. Section 2 discusses previous research which becomes main references for this research. Section 3 discusses methods which are about text preprocessing of the dataset taken from LinkedIn website, and random forest classifier model for job vacancy recommendation. Section 4 presents the experiment, results and discussions, and the last section 5 conveys the conclusion of this paper.

2. LITERATURE REVIEW

The recommendation system in this research is based on two main foundations; random forest classifier and text analysis using word embeddings and cosine similarity. The research entitled "Book Data Content Similarity Detector with Cosine Similarity" was conducted by (Soyusiawaty, 2018). The research has several stages, namely information retrieval, pre-processing, TF-IDF, and cosine similarity. Information retrieval which involves some data to return relevant information to the user. The second stage, pre-processing, is converting the data into a structured format with tokenization, stop word elimination, and stemming. TF-IDF looks for how often a term appears in a document. Cosine similarity is used to measure the similarity of the content to the distance vector. The closer the vector distance, the more equal. System accuracy reaches 82.5% with suggestions for improvement to expand the scope of the book's content.

The classification study with the title "Chrysanthemum Abnormal Petal Type Classification using Random Forest and Over-sampling" conducted by Yuan et al. (2018) carried out flower classification using the Random Forest Classifier algorithm and imbalanced datasets. The results of this study indicate that the Random Forest Classifier algorithm has high accuracy, around 91%-98% with imbalanced datasets. Another classification study entitled "Random Forest for Diabetes Diagnosis" by Benbelkacem (2019) shows that the Random Forest Classifier algorithm has a low error rate of 0.21 with suggestions for adding classification features.

A similar recommendation system research was conducted by Jain (2019) with the title "Job Recommendation System based on Machine Learning and Data Mining Techniques using RESTful API and Android IDE". This study recommends jobs based on the candidate's interest in a company or vice versa. This research uses data mining technology and Vector space model. Vector space model as a TF-IDF representation that works by mapping certain words with documents. This study has a weakness in the form of no measurement of predictive accuracy. Suggestions for improvement from this research is to cooperate with LinkedIn so that it gets data from every user. In addition, classification should be carried out before the recommendation process is carried out.

Another study that raised the topic of a similar recommendation system was carried out by Olowolayemo et al. (2018) with the title "University Based Job Recommender & Alumni System" which uses Data Mining and Jaccard Similarity technology. This application is available from two platforms, namely iOS and Android. The evaluation was carried out by experimenting with 3 iOS users and 7 Android users and no accuracy measurement was used. The research suggestions include collaborating with job search websites and making the user interface more user friendly. Based on the above-mentioned literatures, this study attempts to implement Random Forest Classifier into a job vacancy classification which recommends a vacancy based on skills and professions. Martinez-Gil conducted a study for job offer recommendations using random forest and svm classifier that worked accurately using both methods. However, the study did not explain in more detail about the job offer portal or job category (Martinez-Gil, 2018). In this study, job vacancies portals and categories of classified IT professions are described

3. METHODOLOGY

The job vacancy recommendation is a system designed to recommend job vacancies based on LinkedIn profiles. This recommendation system aims to make it easier for job seekers to find jobs that match their profession. This recommendation system takes into account experience, skills, educational background, previous job titles, and certifications held. In addition, this application program serves to make decisions about choosing a profession at the next career level. The recommendation system is made using an Application Programming Interface (API) that will be used in web-based applications. The recommendation system is equipped with a web-based interface to make it easier for users to view and recommend. This interface has several layouts in the form of input requests for LinkedIn profile links. Based on this input, the system will provide recommendations by classifying jobs and retrieving job vacancies data available on other similar websites. Figure 1 shows the design of the job vacancy recommendation system.

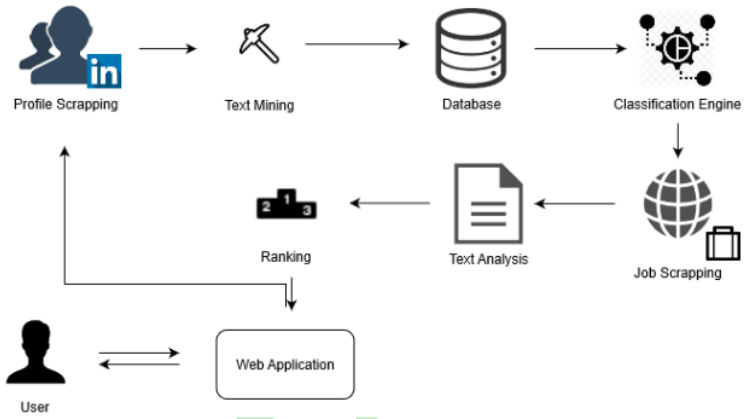


Figure 1 - The design of the proposed system

The design of the recommender system in fig. 1 has an initial process of profile scrapping from the LinkedIn website. Profile scrapping is intended to retrieve data from the web to be processed into datasets. The dataset will be used for the classification process so that it can provide appropriate recommendations. The results of the scrapping profile are processed with text mining technology which is useful for extracting data. Text mining carried out in this process is in the form of feature engineering which functions to form features that are used for classification. In addition, text mining also aims to compile datasets that will be used in the training and testing process.

The dataset that is being used in the classification process is being stored in a database. After that, the classification process is carried out using the Random Forest Classifier algorithm. The classification results will be displayed when the system provides job recommendations. The classification results are also used to find jobs on other websites. This process is made using a web scrapping technique. The job description history on the profile and job description on the job search

website will be compared for similarities using the word embedding method. The level of similarity will be sorted from the highest value and will be recommended to users.

The implementation process is divided into five parts. The first part is the creation of a dataset by capturing links and determining each class label. The second part is taking the classification feature based on the data taken from LinkedIn. The third part is annotating the classification dataset. The fourth part implements job classification using the Random Forest Classifier algorithm. The last part is to carry out the process of implementing a recommendation system using word embedding and cosine similarity methods.

3.1 Dataset Generation

Creating a classification dataset requires target classes. The classification class used is the type of work that is in the Information Technology category (IT). Based on the types of jobs in the IT category, there are 5 job classes that can be quantified based on experience, education, achievement, as well as certification or volunteer activities. Quantification means that ones can calculate how many skills they have if they want to have a certain profession. In addition, quantification can be seen from the number of skill certifications possessed or educational background and work experiences. The types of skills in Table 1 are obtained from several vacancies on job seeker websites. The job classes are: Software Engineer, Mobile Developer, DevOps, Data Scientist, and Data Analyst.

Table 1 – Examples of skills and professions

Skills	Profession Class
Golang	Software Developer
Python	Software Developer
Swift	Mobile Developer
Kotlin	Mobile Developer
Ansible	DevOps
Jenkins	DevOps
Scikit-Learn	Data Scientist
Tensorflow	Data Scientist
Tableu	Data Analyst
PowerBI	Data Analyst

The search for LinkedIn profile links is carried out to collect data on people who have the same profession so that a dataset can be formed. The link profile search carried out is a profile that has one of the experiences of the five classes. The link used to retrieve the data is a link with several experiences of the same work class. In addition, the link used is a link that has a profession of one of the job classes in his last experience. The number of links or links taken is 100 from each job class. The classification process uses five job classes so that the links taken are 500. The links are collected into a csv format file along with the job classes. The process of scraping or retrieving LinkedIn data is divided into automating logins, automating scrolling, and retrieving data attributes. Attribute data is then processed for the feature retrieval process. Feature retrieval is aimed at making classification datasets.

3.2 Feature Extraction

Feature retrieval for classification is done by looking at the scraping data. Scrapping data include work experience, skills, volunteer activities, education, and achievements. Based on these data, six main features can be made which are then broken down into 26 features. Based on work experience data, features such as: how much total work experience is correlated with job class. In addition, a feature is taken in the form of a large number of job titles related to job classes. The average duration of each job is also taken as a classification feature.

A feature that calculates the total work experience in one work class by mapping job titles with job datasets. After that, the duration of the job with the job title is taken and added up. The job dataset is created by creating a csv file containing the job title of the job class. For example, titles such as System Administrator or DevOps Engineer relate to the DevOps job class. After the dataset is created, mapping is carried out. if it matches, then the duration attribute is taken and summed with other similar jobs. Total work experience is expressed in months.

3.3 Dataset Annotation

Annotations or labeling for classification datasets are made by viewing some data such as work. If the last or majority job title comes from the Data Scientist category, the label used is DS, as well as the other four categories. After the annotation, verification is carried out with someone who has one of the five job classes. Dataset verification takes multiple rows from each dataset as a sample. Verification was carried out with fifteen professionals who had these careers and experiences using Google Forms. Fifteen professionals were divided into 5 occupational groups, with a total of 3 people in each work class. In addition, the fifteen people were from several different companies.

In addition, the validation of features for the classification dataset is also carried out. Dataset validation is carried out simultaneously with dataset validation for each job class. Based on the results of the interview, the number of skills, job titles, average duration, and number of certifications have the most influence on job classification. The feature that is considered quite influential is the duration of working in one place. The feature that is considered not too influential on the results of the classification is educational background.

3.4 Random Forest Classification

Random Forest was first introduced by Breiman (2001) is an ensemble approach that can be used to perform classification and regression tasks. The main principle behind the ensemble method is that a set of weak algorithms can combine to form a strong algorithm. Breiman said that a random forest is a combination of tree predictors (for example a decision tree which in the ensemble method is considered a weak algorithm). Each input will be classified by each decision tree in the random forest and the most common results are used as the final classification. Random Forest is widely applied and developed in many research areas (Zhang, 2017).

After making the dataset using web scrapping and feature retrieval methods, a job classification model was created using the Random Forest Classifier algorithm. This study makes five models with varying number of features. The five models have 5, 10, 15, 20, and 26 features. To test the capabilities of the five models, two types of K-Fold Cross Validation were carried out. Table 2 shows the Random Forest model used and the feature numbers.

Table 2 - An example of a table

Model No.	Feature Numbers
1	1,2,3,4,5
2	1,2,3,4,5,6,7,8,9,10
3	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15
4	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20
5	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26

3.5 Word Embedding and Cosine Similarity

Making a recommendation system has several processes that are almost the same as creating a dataset, namely scraping and retrieving features. The distinguishing stage is after the classification process. The classification process produces a job class which is then formed into keywords that will be searched for on job vacancies websites. This study uses two websites for taking job vacancies, namely LinkedIn and Glassdoor. The process of taking job vacancies on both websites requires logging in to get credentials for taking job vacancies.

After taking job vacancies on each website, measurement of the similarity between the job description text and the description contained in the LinkedIn profile is carried out. The similarity measurement process is done by assessing cosine similarity from job vacancies descriptions and linkedin profile descriptions. To convert the description text into vector form so that the cosine similarity calculation process can be carried out, the word embeddings method is used.

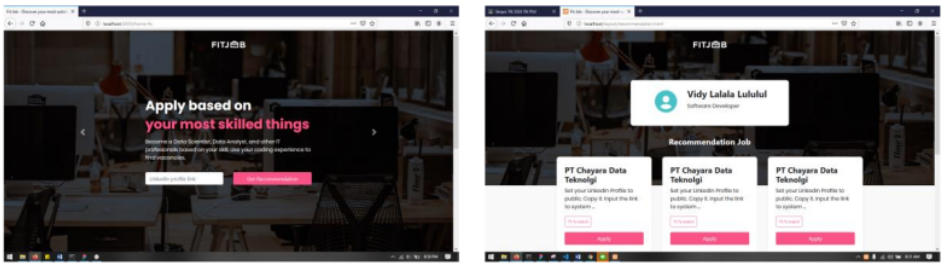
The text analysis process is carried out to measure the similarity between the profile and job description of job vacancies on other websites. This text analysis process is in the form of using word embedding to convert sentences or texts into vector form and using cosine similarity to measure the level of similarity in a sentence.

Word embedding maps words in the vocabulary into vectors that have real values (Bhir, 2017). Word embedding records the semantic and syntactic meaning of words from large, unlabeled corpus of data. In this study, word embedding used is pretrained word embedding with the name FastText using data in the form of wikipedia text data.

Cosine similarity is a measure of the similarity between two vectors by measuring the degree of cosine between the two vectors. The cosine value of the angle between two vectors determines whether the two vectors have the same direction (Shoyusiawati, 2018).

3.6 Web-based Implementation

The user interface in this recommendation system has three layouts. The first layout for input LinkedIn profile links. The second layout to display the results. The third layout is to display the about input. Figure 2 shows the web-based screen capture of the recommender system which ask user to input the LinkedIn profile link (Figure 2(a)) and resulting recommendation jobs (Figure 2(b)).



(a) (b)

Figure 2 - (a) Website main page; (b) Recommendation result page

4. DATA ANALYSIS AND DISCUSSION

Testing and validation for the Random Forest Classifier model was carried out with a total of 500 data using the k-fold cross-validation technique. In addition, several configurations of the number of features and the number of folds were carried out to see the optimal features in forming a classification model. The amount of data is divided into five classes of work with each class totaling 100 data. Table 3 is the results of testing using k-fold cross validation, while Figure 3 depicts the comparison chart of k-fold cross validation results.

Table 3 - K-fold cross validation results

No of Features	5-Fold Accuracy	10 Fold Accuracy
5	0.662	0.664
10	0.934	0.938
15	0.94	0.938
20	0.93	0.93
26	0.934	0.94001

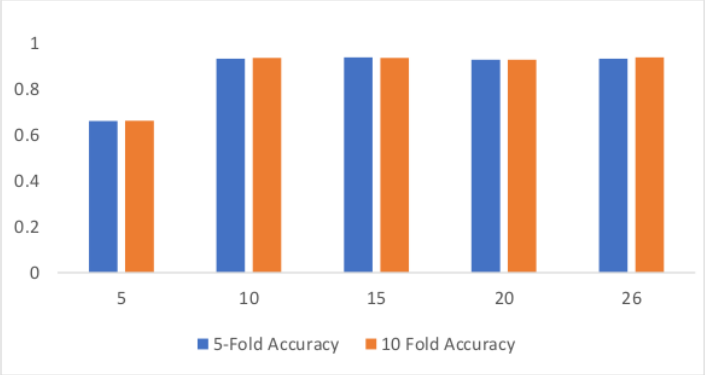


Figure 3 - The comparison of accuracy rate using k-fold cross validations

Based on Table 3, the model with 26 features is used because it has the highest accuracy. In addition, an examination was carried out whether the 26 features were not redundant or had no influence on the classification model. The examination is carried out using feature selection. Feature selection in this study uses the feature importance technique, which is to see the results obtained from the Random Forest Classifier algorithm. The results of feature importance show that all the features used have almost equal weights with each other so that it can be concluded that all features have a uniform effect. Therefore, 26 is used in forming the classification model. The results of the 10-fold cross-validation test using 26 features can be seen in Table 4.

Table 4 - Average classification report on testing

Classification Class	Precision	Recall
Data Analyst	0.954	0.923
Data Scientist	0.94	0.958
DevOps	0.974	0.98
Mobile Developer	0.934	0.939
Software Developer	0.92	0.896

The last test carried out is the hold-out test which is carried out using data testing on the entire system including random forest classifier and text analysis using word embeddings and cosine similarity. The testing data includes 10 LinkedIn profiles data that have never been used in the modeling process, then the results will be evaluated manually by humans to compare the results of recommendations. The results of the hold-out test are listed in Table 5.

Table 5 - Testing results

No of Test	Real Occupation	Recommendation Results
Testing 1	Data Scientist	Data Scientist
Testing 2	Data Scientist	Data Scientist
Testing 3	Devops	Devops
Testing 4	Devops	Devops
Testing 5	Data Analyst	Data Analyst
Testing 6	Data Analyst	Data Analyst
Testing 7	Mobile Developer	Mobile Developer
Testing 8	Mobile Developer	Mobile Developer
Testing 9	Software Developer	Software Developer
Testing 10	Software Developer	Software Developer

Based on Table 5, for the first ten hold-out data, it can be concluded that the classification results were appropriate. Output was called matched when the real job and classification results have the same and match label. In addition, the recommendations given are in accordance with the classification results of the work class. The recommendation results are said to be appropriate if the job classification and job occupation are similar as those of the user. This can be seen from the LinkedIn profile link inputted by the user which is the result of data processing.

In addition, it was found that some job postings had ambiguous or ambiguous job titles and descriptions. This can be seen because there are several links that have the results of the Data Scientist classification that received recommendations similar to data analysts because the two positions are classified as overlapping with each other. In addition, there are several companies that have job postings that do not match the job title and job description.

5. CONCLUSION

This research has produced a recommendation system for five types of work in the IT field, namely: Software Engineer, Mobile Developer, DevOps, Data Scientist, and Data Analyst. Features are taken by analyzing the scraping data. Based on the analysis, features are obtained in the form of calculating skills, job titles, duration, average duration, and certification for each job class. In addition, the latest education is also used as a feature. Annotation or labeling for the formation of the dataset was successfully carried out manually based on the results of the last work. Classification

is done using the Random Forest Classifier algorithm, word embeddings and cosine similarity to calculate the similarity. Model testing during training and testing resulted in an accuracy rate of 94% for the classification model built. The recommendations produced are good but it is hoped that they can be improved with different types of professions.

REFERENCES

Benbelkacem, S. & Atmani, B. (2019). Random Forests for Diabetes Diagnosis. 2019 International Conference on Computer and Information Sciences (ICCIS).

Bitte, R. (2016). Announcing the Ninth Annual 2016 Recruiter Nation Report. https://www.jobvite.com/jobvite-news-and-reports/announcing-the-ninth-annual-2016-recruiter-nation-report/ [Accessed 18 January 2021].

BPS.(2018). *Workforce Condition in Indonesia* https://www.bps.go.id/publication/2018/06/04/b7e6cd40aaea02bb6d89a828/keadaan-angkatan-kerja-di-indonesia-februari-2018.html [Accessed 18 January 2021].

Breiman, L. (2001). Random Forests. Machine Learning, 45, 5-32. http://dx.doi.org/10.1023/A:1010933404324

Fernández-Delgado, M. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? Journal of Machine Learning Research, 15(90): 3133 - 3181

Jain, H., & Kakkar, M. (2019). Job Recommendation System based on Machine Learning and Data Mining Techniques using RESTful API and Android IDE. 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 416-421.

Lim, V. K. G. e. a. (2016). Unemployed and Exhausted? Job-search Fatigue and Reemployment Quality. Journal of Vocational Behavior. Academic Press Inc., Vol. 92.

Martinez-Gil, Jorge & Freudenthaler, Bernhard & Natschläger, Thomas. (2018). Recommendation of Job Offers Using Random Forests and Support Vector Machines.

Olowolayemo, A., Harun, K. & Mantoro, T. (2018). University Based Job Recommender & Alumni System. 2018 International Conference on Computing, Engineering, and Design (ICCED).

Soyusiawaty, D. & Zakaria, Y. (2018). Book Data Content Similarity Detector. Proceeding of 2018 12th International Conference on Telecommunication Systems, Services, and Applications.

Turczynski, B. (2020). HR Statistics: Job Search, Hiring, Recruiting & Interviews. https://zety.com/blog/hr-statistics#online-recruiting-and-social-media-statistics [Accessed 18 January 2021].

Wang, Y. e. a. (2019). Epileptic State Classification for Seizure Prediction with Wavelet Packet Features and Random Forest. Proceedings of the 31st Chinese Control and Decision Conference, p. 3983–3987.

Xu, X. a. Y. F. (2018). Sentences similarity analysis based on word embedding and syntax analysis. International Conference on Communication Technology Proceedings, ICCT, p. 1896–1900.

Yuan, P., Ren, X., Xu, F., Chen, J. (2018). Chrysanthemum Abnormal Petal Type Classification using Random Forest and Over-sampling, IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 275-278, doi: 10.1109/BIBM.2018.8621234.

Zhang Y., Song B., Zhang Y., Chen S. (2017). An Advanced Random Forest Algorithm Targeting the Big Data with Redundant Features. In: Ibrahim S., Choo KK., Yan Z., Pedrycz W. (eds) Algorithms and Architectures for Parallel Processing. ICA3PP 2017. Lecture Notes in Computer Science, vol 10393. Springer, Cham. https://doi.org/10.1007/978-3-319-65482-9_49

IT Job Vacancy Recommender System using Random Forest Classifier

ORIGINALITY REPORT

5%	4%	4%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.researchgate.net Internet Source	2%
2	dokumen.pub Internet Source	1%
3	I Indriyanto, I D Sumitra. "Measuring the Level of Plagiarism of Thesis using Vector Space Model and Cosine Similarity Methods", IOP Conference Series: Materials Science and Engineering, 2019 Publication	1%
4	"Evolutionary Computing and Mobile Sustainable Networks", Springer Science and Business Media LLC, 2021 Publication	1%
5	ebin.pub Internet Source	1%
6	www.tilt.tv Internet Source	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		