Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

RECOMMENDATION SYSTEM FOR ONLINE JOB VACANCY USING MACHINE LEARNING MODELS

YOSUA F SIMANJUNTAK[1], ANTONI WIBOWO[2]

[1]Bina Nusantara University, Computer Science, Jakarta, Indonesia

[2] Bina Nusantara University, Computer Science, Jakarta, Indonesia

E-mail: [1]yosua.simanjuntak@binus.ac.id, [2]anwibowo@binus.edu

ABSTRACT

The Recommendation is used to exploit the relations among known features and content that describe items (content-based filtering) or the overlap of similar users who interacted with or rated the target item (collaborative filtering). To combine these two filtering approaches, current model-based hybrid recommendation systems typically require extensive feature engineering to construct a user profile. Name Entity Recognition (NER) provides a straightforward way to identifies one item from a set of other items that have similar attributes of the related objects. However, due to the large scale of the data used in real world recommendation systems, little research exists on applying NER models to hybrid recommendation systems in job vacancy environment. This paper is proposed a way to adapt the name entity recognition approaches to construct a real hybrid job recommendation system. Furthermore, in order to satisfy a common requirement in recommendation systems the approach of accuracy, precision, recall and F-measure is using in this recommendation system in a principled way. The experimental results demonstrate the efficiency of our proposed approach as well as its improved performance on recommendation precision.

Keywords: *Recommendation System, Job Recommendation, Content-based filtering, Name Entity Recognition, TF-IDF*

1. INTRODUCTION

Job vacancies that can be accessed on the internet help users in finding jobs and candidates. Results indicate that the Internet is an especially valuable job search tool for workers who are distant from the labor market. More and more candidates hunt their jobs from the Internet, bringing an unprecedented increase of human resources information, which leads to the problem of information overload in human resources services [5]. As the amount of information grows, a recommendation system is necessary to help match the right candidate with the right job. To do so, recommendation techniques such as content-based filtering, collaborative filtering and hybrid approaches can be applied [2]. One of the most popular recommender approaches is content-based filtering, which exploits the relations between applied to jobs and similar features among new job opportunities for consideration [4].

The content-based approach matches candidate profiles with employer profiles and job requirements. Initially based on keyword search, content-based filtering was improved into a statistical inference and semantic engine to match profiles rather than keyword [2]. Machine learning is a subfield of computer science that gives computers the ability to learn without being explicitly programmed [1].

Machine Learning uses computers to simulate human learning and allows computers to identify and acquire knowledge from the real world and improve performance of some tasks based on this new knowledge. Machine learning algorithms are being used in recommendation system to provide users with better recommendation. Recommender systems use many various similarity functions to compute similarity between users, between items or between users and items, some similarity functions are heuristic, and others are learnt models from underlying data using machine learning techniques. Performance of various recommendation methods was evaluated by using popular traditional measures in the field of Information Retrieval as well as Recommender Systems such as accuracy, precision, recall, and f-measure. A widely used technique for recognizing

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

entities is named entity recognition (NER). NER refers to identifying all the occurrences belonging to a specific type of entity in the text. NER tasks require a large amount of annotated data that could be extremely cumbersome to produce [23]. More developed NER models can recognize, identify, and classify entities as well. There are many scenarios and use cases of NER technology, like classifying content for news providers, powering content recommendations, customer support, efficient search algorithm, etc. These use cases are domain specific, and the need for identifying words is also different. For example, in classifying content for news providers, the words required to be identified are attack, crime, politics, etc. For the fashion website database search engine, the words required to be identified are clothes, shoes, color, size, etc. For every different requirement, a new NER model can be created, or the existing one can be customized to identify some specific words [24].

This paper proposed to use a machine learning model for supporting the job recommendation system. However, to make a better prediction model in the recommendation system, content-based must be combined as a hybrid system to avoid cold start problems. In the same way, machine learning model such as SVM, KNN, and Naïve bayes are used to develop classification algorithm with the help of combination TF-IDF and NER in preprocessing task to increase the performance in recommending job. On the other hand, measuring the model is another challenge in this thesis. Thus, the traditional measure such as accuracy, precision, recall and f-measure will be used to measure the performance of the model. In conclusion, an overview of comparing content-based filtering and the hybrid system will be provided as an evaluation to find the best method to recommend jobs and candidates.

2. RELATED WORKS

Recently, job seeking and recruiting websites have been experiencing a striking rise. As the amount of information grows, a recommendation system is necessary to help match the right candidate with the right job. To do so, recommendation techniques such as content-based filtering and hybrid approaches can be applied. The content-based approach matches candidate profiles with employer profiles and job requirements. Initially based on keyword search, content- based filtering was improved into a statistical inference and semantic engine to match profiles rather than keyword. Besides, previous studies state that the challenge of matching candidates and jobs is grounded in the interactionist theory of behavior and believe that interactions are important for recommendation as they strongly influence the candidate's job choice and employer's hiring decision. Some interaction-based recommendation systems, such as CASPER, make use of collaborative filtering to recommend jobs to users based on what similar users have previously liked. Hybrid systems are also exploited to match people and jobs. A hybrid system proposed by Malinowski is based on the idea that a good match between people and jobs needs to consider both the preferences of the recruiters and the candidates. As a result, the matches between jobs and candidates are predicted according to candidate CVs and employer descriptions and requirements as well as previous rating information. The limitation of this method is that it is difficult for the users to rate the jobs they have not worked on yet.

3. RESEARCH FRAMEWORK

The job recommendation system helps both candidates to find jobs and the recruiter to find the match candidate. However, finding a suitable job and candidate is time-consuming and inefficient due to information overload. One of the major problems in the recommendation system is the cold start problem where there is no historical data for the new user to be recommended from the job list. In addition, finding the best method to improve the performance recommendation is one of the major processes in the research. Similarly, evaluation needs to be done to get best clustering algorithm in providing recommendation.

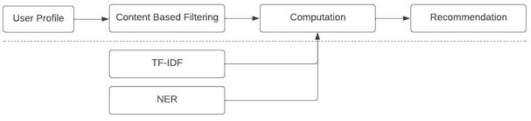

Figure 1: Diagram of propose method

This research is proposed to create a job recommendation system to speed up the process and to increase the efficiency of finding jobs and candidates, improve the performance of the recommendation, and measure the machine learning models to find the best algorithm to recommend the job and candidate to the user. This research used Python programming language version 3.8 and with the support of python library such as Natural Language Toolkit (NLTK) for

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

providing function to preprocessing the text using TF-IDF and NER. In the same way, Python will be used to create the algorithm for content based, collaborative and hybrid of the combination content and collaborative. Equally, scikit-learn library is provided for creating SVM, KNN and Naive Bayes machine learning model as well as split the data to training and testing. In the process of the evaluation, this research will be used cofunction matrix measure which available in the scikit-learn library.

The methodology is divided into nine stages about the process of the research. Below are steps:

A. Identification Problem

First of all, before designing the recommendation system, the problem must be identified in order to create gather all the requirements for creating the system. The problems in the research are finding the best algorithm, combine the content-based and collaborative filtering to solve the cold-start and sparsity problem, and measure the model with the evaluation method.

B. Study Literature

The study literature stages are the main important step in the research. This stage is to determine the method that will be used in the research be carried out. Gather the journal is needed to find the weakness and strengths related to the topic of the job recommendation system. To conclude, the method of the problem and the new idea is coming from the journal could increase the performance of the system.

C. Collecting Data

In order to create the job recommendation system, the dataset must be collected to create the machine learning model. The dataset used in the research was collected from kaggle.com. The datasets are about 24.475 job postings, and 200 candidates resume.

D. Preprocessing Data

1.Remove Punctuation

Firstly, the punctuation marks such as comma, special character, semicolon, colon, quantitation marks, dash, hyphen, bracers, parentheses, etc. will be removed from the text as these could consist of unimportant weight. It can be achieved by using the Python String Punctuation function.

2. Case Folding

A sentence consists of a combination of both uppercase and lowercase letters. Thus, the process of case folding is to reduce all the text to lower case. The functions lower() is available in python for converting the text.

3. Tokenization

This step is to break the word with the whitespace as the separators and storing each of the words in lists called as token. This can be done using NLTK tokenize in python.

4. Remove Stopwords

The stopwords such as a, an, the, in, etc. are often added to the sentence to make it grammatically correct. Those words must be removed to give clarity to the model and focusing only on the word which defines the meaning of the text. NLTK Python library can be used to remove the stopwords.

5. Stemming

In the job recommendation, this step is important especially in the job posting and candidate experience. The stemming process reduced the inflected word to their root form. This process helps in standardizing the text and removing the number of tokens. NLTK Porter Stemmer function is used in this process to stem the text.

6. Vectorization

The vectorization encoded the final tokens into numbers to create feature vectors in order to give understanding to the algorithm. One of the types of vectorizations namely the TF-IDF algorithm used in this step. The algorithm stored the weightage of each word by calculating the frequency the word occurs in a document using a formula.

7. NER

The NER method will divide the data into entity or segmentation level. The job description text is then classified into named entities based on their experience. Then every detected entity is classified into a predetermined category such as person, organization, time, location etc.

E. Content Based, Collaborative and Hybrid

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The content-based and collaborative filtering alone method will proceed, and the output of that method will be evaluated by similarities. On the other hand, the hybrid method will use the output from the content based filtering as the source for collaborative filtering for the user does not have historical data in applying for a job.

F. Classification Model

After the splitting process finished. The next step developed classification machine learning model such as SVM, KNN, and Naïve Bayes in the job recommendation system. The machine learning model worked with the training data to identify the classification in the data with the similar types. In addition, the machine learning model will be evaluated by confusion matrix to find the best algorithm for recommending the data.

G. Training and Testing Model

After the preprocessing step is done, the machine learning model need to be trained from the dataset collected. Therefore, the process of splitting data into training and testing is crucial. The training process used 70% of the data in order to create the machine learning model. The training data consist of the job description, traffic view and both candidate's experience and interest to the job. Equally, the testing used 30% of the dataset to evaluate the accuracy of the model.

H. Evaluating Model

The job recommendation system applied confusion matrix as the evaluation of the research. The confusion matrix handled the accuracy, precision, recall, and f-measure to get the performance for machine learning model. Content-based and hybrid recommendation system will be evaluated based on each of machine learning model.

I. Report

In the report stage, the result of the evaluation model will be documented in this stage. This is the last part of the thesis which will be inserted into the report research. In another word, the report can be used for another research purpose analysis as a reference

4. THEORY AND METHODS

A. Recommendation System

Recommender Systems are software tools and techniques providing suggestions for items to be of use to a user. The suggestions provided are aimed at supporting their users in various decision-making processes, such as what items to buy, what music to listen, or what news to read. Recommender systems have proven to be valuable means for online users to cope with the information overload and have become one of the most powerful and popular tools in electronic commerce [13]. Therefore, recommendation analysis is often based on the previous interaction between users and items, because past interests and proclivities are often good indicators of future choices [14].

B. Collaborative Filtering

The main idea of collaborative recommendation approaches is to exploit information about the past behavior or the opinions of an existing user community for predicting which items the current user of the system will most probably like or be interested in [15]. Collaborative filtering methods produce user specific recommendations of items based on patterns of ratings or usage (e.g., purchases) without need for exogenous information about either items or users [13].

C. Content Based Filtering

Content-based recommendation systems try to recommend items similar to those a given user has liked in the past, whereas systems designed according to the collaborative recommendation paradigm identify users whose preferences are similar to those of the given user and recommend items they have liked [13]. Content based systems have certain basic components, which remain invariant across different instantiations of such systems. In most cases, it is preferred to convert the item descriptions into keywords. Therefore, content-based systems largely, but not exclusively, operate in the text domain. Many natural applications of content-based systems are also text-centric [14].

D. Cosine Similarity To find similar items, a similarity measure must be defined. In item-based recommendation approaches, cosine similarity is established as the standard metric, as it has been shown that it produces the most accurate results. The metric measures the similarity between two n-dimensional vectors based on the angle between them [15].

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$sim(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{|\vec{a}| * |\vec{b}|} \quad (1)$$

Figure 2: Cosine similarity equation

D. Named Entity Recognition

A named entity is a word or a phrase that clearly identifies one item from a set of other items that have similar attributes. Examples of named entities are organization, person, and location names in general domain, gene, protein, drug and disease names in biomedical domain [19]. Named entity recognition (NER) is the problem of locating and categorizing important nouns and proper nouns in a text. For example, in news stories names of persons, organizations and locations are typically important. Named entity recognition plays an important role in applications such as Information Extraction, Question Answering and Machine Translation.

Representation	Example		
SGML	<PER>Dr. Doull</PER> from the <ORG>Royal College of Paediatrics</ORG> in <LOC>Wales</LOC> backed the <MIS>Fresh Start</MIS>.		
	Token	**BIO**	**BIOLU**
BIO & BIOLU	Dr.	B-PER	B-PER
	Doull	I-PER	L-PER
	from	O	O
	the	O	O
	Royal	B-ORG	B-ORG
	College	I-ORG	I-ORG
	of	I-ORG	I-ORG
	Paediatrics	I-ORG	L-ORG
	in	O	O
	Wales	B-LOC	U-LOC
	backed	O	O
	the	O	O
	Fresh	B-MIS	B-MIS
	Start	I-MIS	L-MIS
	.	O	O

Figure 3: Name Entity Recognition Example Entity

Named entity recognition plays an important role in applications such as Information Extraction, Question Answering and Machine Translation. Named entity recognition systems are evaluated by running them on human-labeled data and comparing their results against this gold-standard. The comparison is usually at the phrase level, giving full credit for complete boundary and category matches and no credit for partial matches. The commonly used evaluation metrics are the precision and recall which have been borrowed from Information Retrieval evaluation [18].

D. Machine Learning

The goal of a Support Vector Machine is to find a linear hyperplane or decision boundary that separates the data in such a way that the margin is maximized.

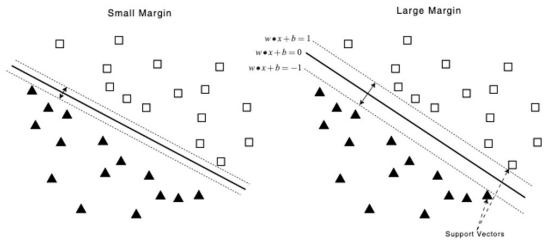

Figure 4: Support Vector Machine Model

The two class separation problem in two dimensions, there are many possible boundary lines to separate the two classes. Each boundary has an associated margin. The rationale behind SVM model is to avoid the misclassify unknown items in the future [13].

5. PROPOSED METHODS

The proposed recommendation system in the research are content-based filtering and hybrid recommendation. The recommendation system was developed using a preprocessing method with the combination of the TF-IDF and NER to improve the performance in recommending job. Machine learning in the research applied SVM, KNN, and Naive Bayes to create the classification model in the data with the same similarity types. To summarize, the machine learning model will be assessed by the evaluation method to discover the best model in providing jobs and candidates.

The job recommendation system applied confusion matrix as the evaluation of the research. The confusion matrix handled the accuracy, precision, recall, and f-measure to get the performance for machine learning model. Content-based and hybrid recommendation system will be evaluated based on each of machine learning model.

6. RESULT AND DISCUSSION

The recommendation used confusion matrix for NER model in preprocessing task as the evaluation of the research. This model is used resume to build the training and testing data. This model is created from previous researcher [26] and improved from this research to get more accuracy.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Evaluation Name Entity Recognition Model

Entity	Precision (%)	Recall (%)	F1 (%)
College Name	74.0	78.0	76.0
Companies worked at	73.0	27.0	39.0
Degree	91.0	88.0	89.0
Designation	67.0	69.0	50.0
Email Address	100.0	86.0	92.0
Graduation Year	41.0	59.0	48.0
Location	0.00	0.00	0.00
Name	95.0	91.0	93.0
Skills	75.0	46.0	57.0
Year Of Experience	0.00	0.00	0.00

In Summary, the highest score is dominated in degree, skill, designation, name, and college name with the threshold up to 70% of recall. Thus, the entities will be used for predicting the job recommendation.

Table 2: Evaluation of Accuracy (Acc), Precision (Prec) and Recall (Rec) for job recommendation model

Dataset	Model		Acc (%)	Prec (%)	Rec (%)
Job Desc + Resume	CBF - TFIDF - NER	**KNN**	**71.0**	**70.0**	**71.0**
		SVM	69.0	71.0	69.0
		NB	63.0	67.0	63.0
Resume	CBF - NER - TFIDF	KNN	77.7	77.7	77.7
		SVM	**83.7**	**73.0**	**84.0**
		NB	81.4	71.0	81.0
Job Desc	CBF - NER - TFIDF	KNN	74.4	55.0	74.0
		SVM	**76.7**	**59.0**	**76.0**
		NB	74.4	75.0	74.0

In this table, the evaluation model shows the highest score of machine learning model is KNN which is 71% of accuracy, 70% of precision and 71% of recall using dataset both job description and candidate resume. On other hand, the highest score using resume dataset is CBF, NER and TFI-DF model using SVM linear machine learning model which is 83,7 of accuracy, 73% of precision, and 84% of recall. Same as the last dataset which is job description having the highest score of SVM model. In summary, the hybrid model CBF, TF-IDF, and NER are giving the best output over 70% in average for KNN and SVM machine learning model.

7. CONCLUSION AND FUTURE WORKS

In the future, this model can be used to other model such as collaborative filtering recommendation and can be applied with the other deep learning model. Also, the model can be improved with the adding of number of resumes to increase model in the training so the accuracy could be increased.

REFERENCES:

[1] Theobald, O. (2017). Machine Learning For Absolute Beginners.

[2] Yao Lu, S. E. (2013). A Recommender System for Job Seeking and Recruiting Website. Proceedings of the 22nd international conference on World Wide Web companion (pp. 963-966). ResearchGate.

[3] Yang, S., & Korayem, M. (2017). Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive Statistical Relational Learning approach. ELSEVIER, 1-9.

[4] Shuo Yang, ∗. M. (2017). Combining content-based and collaborative filtering for job recommendation system: A cost-sensitive Statistical Relational Learning approach. ELSEIVER.

[5] Weijian Chen, X. Z. (2017). Hybrid Deep Collaborative Filtering for Job Recommendation. IEEE, 275-280.

[6] Shivam Bansal, A. S. (2017). Topic Modeling Driven Content Based Jobs Recommendation Engine for Recruitment Industry. th International Conference on Information Technology Selection and/or peer-review under responsibility of the organizers of ITQM 2017 and Quantitative Management, ITQM 2017 (pp. 865-872). India: Elsevier B.V.

[7] Bharat Patel, V. K. (2017). CaPaR: A Career Path Recommendation Framework. 2017 IEEE Third International Conference on Big Data Computing Service and Applications (BigDataService) (pp. 23-30). California: IEEE.

[8] Anika Gupta, D. D. (2014). Applying Data Mining Techniques in Job Recommender System for Considering Candidate Job Preferences. 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1458-1465). Delhi: IEEE.

[9] Pradeep Kumar Roy, ∗. S. (2020). A Machine Learning approach for automation of Resume Recommendation system Recommendation system. International Conference on Computational Intelligence and Data Science

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

(ICCIDS 2019) (pp. 2318-2327). Vellore: Elsevier B.V.

[10] Minh-Luan Tran, A.-T. N.-D. (2017). A Comparison Study for Job Recommendation. KICS-IEEE International Conference on Information and Communications with Samsung LTE & 5G Special Workshop (pp. 199-204). Phu Yen: IEEE.

[11] Mamadou Diaby, E. V. (2013). Toward the Next Generation of Recruitment Tools: An Online Social Network-based Job Recommender System. 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 821-828). Paris: IEEE.

[12] Yingya Zhang, C. Y. (2014). A Research of Job Recommendation System Based on Collaborative Filtering. 2014 Seventh International Symposium on Computational Intelligence and Design (pp. 533-538). Beijing: IEEE.

[13] Ricci, F., Rokach, L., Shapira, B., & Kantor, P. B. (2011). Recommender Systems Handbook. New York: Springer.

[14] Aggarwal, C. C. (2016). Recommender Systems: The Textbook. New York: Springer.

[15] Dietmar Jannach, M. Z. (2011). Recommender Systems An Introduction. Cambridge: Cambridge University Press.

[16] Bruno Trstenjaka, S. M. (2014). KNN with TF-IDF Based Framework for Text Categorization. 24th DAAAM International Symposium on Intelligent Manufacturing and Automation. Vienna: ELSEVIER.

[17] B.Upendraa, D. A. (2016). KNN TFIDF Based Named Entity Recognition. IJSDR, 35-39.

[18] Zitouni, I. (2014). Natural Language Processing of Semitic Languages. Redmond: Springer.

[19] Jing Li, A. S. (2020). A Survey on Deep Learning for Named Entity Recognition. IEEE.

[20] Mykhailo Granik, V. M. (2017). Fake News Detection Using Naive Bayes Classifier. 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (pp. 900-903). Vinnytsia: IEEE. [21] URL Date Stamp Time Stamp GMT and dd/mm/yyyy

[21] Herley Shaori Al-Ash, W. C. (2018). Fake News Identification Characteristics Using Named Entity Recognition and Phrase Detection. 10th International Conference on Information Technology and Electrical Engineering (pp. 12-17). Depok: International Conference on Information Technology and Electrical Engineering (ICITEE).

[22] Yao Chena, C. Z. (2019). Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training. ELSEVIER.

[23] Bodhvi Gaur, G. S. (2020). Semi-supervised deep learning based named entity recognition model to parse education section of resumes. Springer, 5706-5718.

[24] Hemlata Shelar, G. K. (2020). Named Entity Recognition Approaches and Their Comparison for Custom NER Mode. Taylor & Francis open access journals and publishing, 1-13.

[25] Flach, P. (2012). Machine Learning The Art and Science of Algorithms that Make Sense of Data. Cambridge: Cambridge University Press.

[26] Dev, "Named Entity Recognition With Bert" https://www.kaggle.com/code/sameerdev7/named-entity-recognition-with-bert res (created Apr. 10, 2020).