JURNAL UPI

Journal of Computers for Society

Journal homepage: https://ejournal.upi.edu/index.php/JCS

Journal of Computers for Society

Implementation of Inverse Document Frequency (TF-IDF) and Cosine Similarity Terms in Determining Research Reviewers for Indonesian Education University Lecturers

Hazmi Ramadhan Adli[1,*], *Munir*[2], *Rani Megasari*[3]

[1, 2, 3] Department of Computer Science Education, Universitas Pendidikan Indonesia, Indonesia
*Correspondence: E-mail: hazmi.ramadhan.adli@student.upi.edu

ABSTRACT

Research is a mandatory activity for lecturers at the Indonesian Education University. The Institute for Research and Community Service (LPPM) oversees these research activities. Before research can commence, the research proposal must be tested or reviewed by an Examining Lecturer (reviewer). Reviewers are selected based on the similarity between the researcher's and the reviewer's variables and the availability of the reviewer's quota. This selection process utilizes the Term Frequency Inverse Document Frequency (TF-IDF) and Cosine Similarity methods to measure the similarity between queries and documents, specifically abstracts and scientific fields. This approach results in reviewer recommendations with an accuracy of 83.3%, as validated by experts.

ARTICLE INFO

Article History:
Submitted/Received 20 Feb 2023
First Revised 18 Apr 2023
Accepted 12 Jun 2023
First Available Online 13 Jun 2023
Publication Date 15 Sep 2023

Keyword:
Cosine Similarity,
Research,
Reviewer,
TF-IDF.

1. INTRODUCTION

Lecturers at the Indonesian University of Education are required to engage in research activities. These research and community service efforts must be completed within one calendar year. As mandated by Law Number 14 of 2005 on Teachers and Lecturers, lecturers are recognized as professional educators and scientists, whose primary duties include the transformation, development, and dissemination of science, technology, and art through education, research, and community service (Darmawan, 2020).

Currently, research management at the Indonesian University of Education (UPI) is running using information technology. It cannot be denied that information technology has changed the paradigm and procedures for managing an activity to become more effective and efficient (Indrayani, 2012). With information technology, data can be managed easily, quickly and accurately thanks to sophisticated computers (Eskak, 2020). Starting from the proposal submission stage to reporting the results of research activities and community service. Management of research and community service at the Indonesian University of Education is managed by the Institute for Research and Community Service (LPPM).

LPPM has the task of planning, implementing, developing and evaluating research and community service activities. Overall, the process of planning, implementing, developing and evaluating research and community service activities has run optimally. In the implementation stage there are several processes carried out starting from submitting proposals, reviewing proposals by reviewers, announcing passed proposals, disbursing research and service funds as well as inputting progress reports and final reports.

During the review process it takes a very long time to pair researchers with reviewers who are appropriate to their scientific field because the pairing activity is carried out by the operator manually, not through an automatic system. Pairing reviewers with Research Proposals by considering several things, including similarities in areas of expertise and knowledge between the Proposing Lecturer and Reviewer, then the relationship between the proposed research title and the portfolio, interests and experience of research titles that have been reviewed by the Reviewer. In general, matching this profile is to get a small gap value which has a greater chance of being recommended (Badrul & Utami, 2022).

Another problem is the limited number of reviewer lecturers in the Research Proposal testing process and also the maximum quota that reviewers have to test a number of Research Proposals. The limited number of reviewer lecturers is because to become a reviewer there are certain qualifications and must have a certificate. Then, as a result of the limited number of reviewer lecturers, this has an impact on several research fields where the number of reviewer lecturers is small. So, the pairing of reviewers with Proposal Proposers is based on profile similarity and testing quota availability. The profile similarity in question includes the criteria that are a prerequisite for pairing reviewers with lecturers proposing proposals, namely areas of expertise, knowledge and other parameters that will be examined in this research. Apart from that, when the quota for testing a proposal has been fulfilled, the proposal will be directed to other lecturers by considering the closest profile, competency and interest in certain research according to the proposed proposal.

This pairing of reviewers with proposal proposers will occur many times. Because in this process many possibilities will occur, especially in terms of similar quota availability profiles for testing proposals. So automation is needed in this installation, so that you don't waste a long time in the process of pairing the reviewer with the proposer of this proposal.

There is several research related to this research, including the Implementation of Cosine Similarity Matching in Determining Final Assignment Supervisors, implementing the Cosine

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088

Similarity method in determining final assignment supervisors in order to obtain an optimal guidance process. Cosine Similarity is a method for calculating the similarity (level of similarity) between two objects (Sidorov *et al.*, 2014). In this study, the level of similarity between the title, topic and abstract of students' final assignments was calculated compared with data from the supervisor in the form of the supervisor's expertise and final assignments that had been supervised by the lecturer. Then the Cosine Similarity method will calculate the level of similarity of the two queries. The highest similarity score will appear as the recommended supervisor (Yasni *et al.*, 2018). In research by Zaware, et al entitled "An Effectual Approach For Calculating Cosine Similarity", they say that text similarity plays an important role in text mining. This research suggests a new approach that allows us to calculate the cosine similarity coefficient in a more productive and cost-effective way. It provides many advantages over traditional approaches of calculating similarity and can therefore be applied in a wide range of applications (Zaware, *et al.*, 2015).

Then the research titled "Designing and Making Scholarship Information Search Applications Using Cosine Similarity" explores the use of an information retrieval (IR) based search system to find scholarship information online. This IR system is developed using the vector space model (VSM). To gather scholarship data, a Web Crawler is utilized, specifically the Vietspider Web Crawler, and the collected data is stored in a database. The similarity between scholarship data is determined using cosine similarity, which helps in presenting relevant scholarship information to users based on their search queries (Kurniawan et al., 2014). In another study titled "Text Mining: Text Similarity Measure For News Articles Based On String Based Approach," the Cosine Similarity algorithm is employed to measure the similarity between news articles. This method involves three crucial preprocessing steps: Stop Word Removal, Extracting the Noun, and TF-IDF on news article datasets (Kohila & Arunesh, 2016).

In the research entitled Essay Type Exam Assessment Using the Text Similarity Method, assessment using text similarity with the tf-idf method produces output that matches the user's specifications, but requires quite a long computing time when the data being processed (document text) is large, the system text similarity is not significantly different compared to expert-based assessments, but care must be taken when selecting answers that will be used as keys so that they do not become words that do not contain meaning (bias) (Sulistyo *et al.*, 2015).

Bambang Kurniawan and colleagues conducted a study titled "Classification of News Content Using the Text Mining Method." This research focused on the automatic classification of news articles on news portals. The approach employed in the study was the text mining method, a form of data mining aimed at discovering intriguing patterns within large sets of textual data. The algorithm applied for this classification task was the naïve Bayes classifier, which aids in the classification process. The outcome of the research was a web-based news classification system developed using the PHP programming language and a MySQL database. The study demonstrated that news articles could be fully automatically classified (Kurniawan et al., 2012).

The next research titled "A Text Categorization Method using Extended Vector Space Model by Frequent Term Sets" explores text categorization. It introduces a new limitation of AD-Sup to extract discriminative features from frequently used term sets for classification tasks. The classification results on the Reuters-21578 and WebKB corpora demonstrate that AD-Sup constraints effectively extract useful features, and the combination strategy improves the feature space and classification performance (Yuan et al., 2013).

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2774-7891

Additionally, the study by Riki Ruli et al., titled "Application for Determining Thesis Examining Lecturers Using the Tf-Idf Method and Vector Space Model," applies text mining, TF-IDF, and Vector Space Model (VSM) to develop a system for recommending thesis examiners. Text mining processes the thesis titles and abstracts, while VSM classifies competencies. This system can recommend three lecturers as examiners based on the match between the title, abstract, and classification. The research employs the CRISP-DM software development model, which includes phases such as business understanding, data understanding, data processing, modeling, evaluation, and deployment. The study achieves an accuracy of 93.22% (Siregar et al., 2017).

2. METHODS

2.1. Data Collection

Lecturer reviewer data and Lecturer Research Proposal proposals used in this research were obtained directly from LPPM Indonesian Education University. This data is data from 2019 research proposals, which were selected randomly. The data obtained consists of proposed research proposals, reviewer lecturers.

2.2. Reviewer Determination System Model Flow

In this system, there is a function to accept lecturers' research proposals. From these results, the abstract and scientific field of the proposing lecturer are obtained which have been stored in the database. The abstract and scientific field of the proposing lecturer are processed using text preprocessing to produce basic words. After that, weighting is carried out on each word in the abstract, the proposer's research field and the reviewer's research field. Then an assessment is carried out using cosine similarity. This assessment will result in a value ranging from 0-1. The higher the value, the higher the similarity between the proposal data and the reviewer data. The value that has been obtained is then converted into a percentage by multiplying by 100%, so that the data obtained is then displayed as a result for reviewer recommendations that can be selected by the admin.

2.3. Test Data

The data used in this research, namely lecturer and reviewer research proposals, was obtained from the 2019 lecturer proposal data consisting of the name of the proposer, the scientific field of the proposer, and the abstract of the proposal and will be compared with the reviewer data which consists of the name of the reviewer and the scientific field obtained. from the LPPM UPI Litabmas database.

2.4. Text Preprocessing Stage

Text preprocessing is the initial stage in processing the lecturer's proposal abstract and the reviewer's scientific field data. In this stage there are several steps that will be taken, which can be seen in Figure 1 (Mariel *et al.*, 2017).

Figure 1. Text preprocessing stage.

In this case folding process, the case is equalized to the abstract and the reviewer's scientific field which will be processed into lower case letters. The tokenizing stage is the process of removing punctuation characters and numbers in words from a document.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088

Removed characters like '~!@#$%^&*()_+={}[]\|;:'",./?. The stopwords stage is the removal of words that are considered meaningless or do not have the power to describe something strong, such as conjunctions. The stemming stage is the stage of changing words that have gone through case folding, tokenizing, stopwords removal into base words.

2.5. Term Weighting Stage

After going through the text pre-processing stage which produces only basic word data, the next stage is word weighting. In this stage, the research proposal abstract is referred to as a query and the reviewer's scientific field is referred to as a document. It is hoped that the results of the text similarity calculation will produce a good document ranking.

To carry out word weighting, the first step is to form a word dictionary. Before the process of determining the reviewer, a dictionary is first created. . Example of a dictionary of words in the form of unigrams for the process of determining research proposal reviewers taken from the abstract as a query and the reviewer's scientific field as a document.

Next is the TF and IDF calculation stage. Tf-Idf is a calculation that describes how important words (terms) are in a document.

The final stage is the stage of calculating the weight of each document. At this stage, the weight of the document for words or the weight of the key for the document will be calculated using the formula below (equation 1):

$$w = tf \times idf \tag{1}$$

2.6. Cosine Similarity Stage

Calculate the similarity of the query [document] vector Q with each existing document. Similarities between documents can use cosine similarity. The formula is as follows (equation 2):

$$similarity = \cos(\theta) = \frac{A \,.B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_1 B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2}$$

2.7. Software Development

In building software, the first stage is needs analysis. The software built is a system for determining Lecturer Examiners (reviewers) for Lecturer research proposals. This tool will provide recommendations for lecturers who can become examiners (reviewers) in testing UPI lecturers' research proposals.

The second stage is software design. In this section the researcher will discuss operational process architecture and software interface design. The first process is the operational process of the software, including input and output produced by the system. The input to the search recommendation system for examining lecturers (reviewers) is the abstract of the lecturer's research proposal and the lecturer's scientific field. The output produced by this system is the name of the reviewer lecturer with the match percentage based on the Cosine Similarity calculation method. So that the names that appear can be selected to become reviewers. The second process is interface design. The display of the software displays a list of lecturer research proposals that have been input by the proposing lecturers with a button

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2774-7891

to search for reviewers. Then the output of the reviewer's recommendation list displays with the percentage of similarity resulting from the Cosine Similarity method calculation process.

After carrying out the software analysis and design stage, the next step is for the author to implement it according to the results of the design developed at the software design stage. This implementation process produces functions that can be executed to complete the processes in this software.

The final stage is software testing. At this testing stage, the author uses a black box method to ensure that all functions run well. Black box testing is referred to as behavioral testing. Where the interior structure, logic of the software under test is unknown to the tester. Testing is based on requirements specifications and does not require code analysis. Black box testing is carried out from the end user's perspective. (Praniffa *et al.*, 2023). Testing using the black box method was carried out by the author by paying attention to each function that had been created.

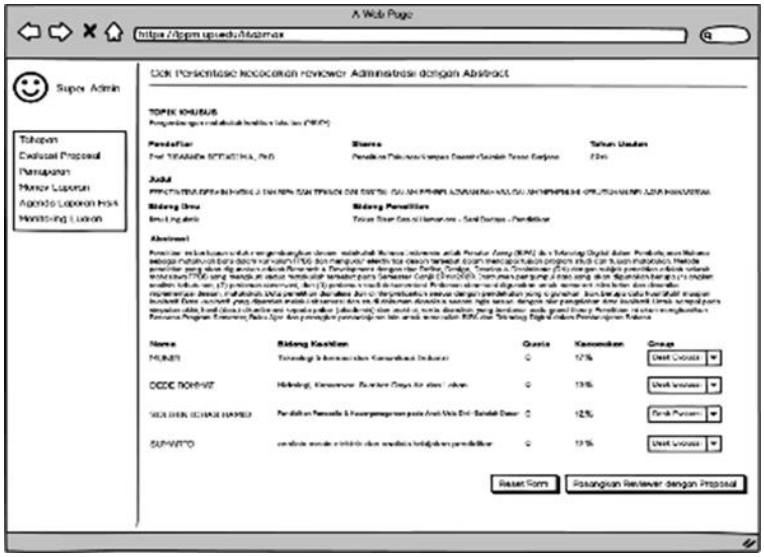

Figure 2. Interface Design.

2.8. Experimental Design

In conducting the experiment, the author used 30 test data originating from 2019 research data which is the focus of this research. This data will be input for the software being built.

Then, the author evaluates the output results of the software to validate whether the recommendations of the examining lecturer (reviewer) can accommodate the proposed abstract. The evaluation carried out was in the form of an expert judgment on Dr. Yadi Ruyadi, M.Sc. as Secretary of LPPM UPI for the 2015-2020 period and Pipin Firdaus, S.Kom. as Research Manager of LPPM UPI.

Then to analyze the results, researchers used a rating scale with the following equation [14]:

$$P = \frac{evaluator's\ assessment\ score}{ideal\ score} \times 100\% \quad (3)$$

Then the calculation results from the rating scale method will be categorized into five categories using the scale in **Table 1**.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088

Table 1. Rating Scale

Percentage Score	Interpretation
P ≤ 20%	Very Less
20% < P ≤ 40%	Less
40% < P ≤ 60%	Enough
60% < P ≤ 80%	Good
80% < P ≤ 100%	Very Good

3. RESULTS AND DISCUSSION

The experimental results are in the form of 30 test data and the results of the recommendations of the Examining Lecturer (reviewer). The results of this test data will then be carried out by expert judgment to evaluate the software output results.

To evaluate software output results. Judgment is carried out by two experts. In carrying out this expert judgment, the experts are asked to evaluate the recommendations of the examining lecturers (reviewers) that appear which can be paired with the abstract of the proposed research proposal (see **Table 2**).

Table 2. Expert Evaluation Results.

Proposer	Suitability	
	Expert 1	Expert 2
1	Yes	Yes
2	Yes	Yes
3	Yes	Yes
4	Yes	Yes
5	No	No
6	Yes	Yes
7	No	No
8	Yes	Yes
9	Yes	Yes
10	No	No
11	Yes	Yes
12	Yes	Yes
13	Yes	Yes
14	Yes	Yes
15	Yes	Yes
16	Yes	No
17	Yes	Yes
18	Yes	Yes
19	Yes	No
20	Yes	Yes
21	No	No
22	Yes	Yes
23	No	Yes
24	Yes	No
25	No	No
26	Yes	Yes
27	Yes	Yes
28	Yes	Yes
29	Yes	No
30	Yes	Yes

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2774-7891

After that, the researcher validates the two experts, if one of the experts answers "Yes" then it is considered an abstract and the recommendation of the Examining Lecturer (reviewer) is considered appropriate. The following are the results of the expert judgment evaluation.

Based on **Table 3**, according to the expert the test data which is considered to be in accordance with the output of the Examiner Lecturer's recommendation (reviewer) and the abstract is 83.3% based on the Yes & Yes answers of 66.7% added to the Yes & No answers of 16.7%. Meanwhile, according to experts, 16.7% did not match the results of the Examining Lecturer's (reviewer's) recommendations with the abstract. If these results are categorized based on the rating scale in **Table 1**, the results are categorized as very good because the results of the Examiner Lecturer's recommendation (reviewer) with the abstract reached 83.3%.

At this analysis stage, the researcher found that when the expert saw an abstract that matched the reviewer's expertise, the expert would conclude that the research proposal could be paired with the reviewer's recommendations produced by the software that the researcher had built. Therefore, the TF-IDF and Cosine Similarity methods can be applied in determining reviewers using abstract data and scientific fields because they get good efficiency values (see **Table 4**).

Table 3. Percentage of Conformity to Expert Judgment Results.

Suitability of Abstract and Recommendations	Percentage (%)
Yes & Yes	66,7%
Yes & No	16,7%
No & No	16,7%

Table 4. Conclusion of Experiment Results.

Suitability of Abstract and Recommendations	Percentage (%)
Yes	83,3%
No	16,7%

4. CONCLUSION

Research to determine reviewers for research proposals for lecturers at the Indonesian University of Education using Tf-IDF and Cosine Similarity produced several conclusions. These conclusions include the following:

(i) In determining the research proposal reviewer, input data is in the form of the applicant's scientific field, research abstract and reviewer's scientific field. The data is processed using preprocessing which includes case folding, tokenizing, stopword removal and stemming. There is a calculation of the comparison of proposer and reviewer variables using the TF-IDF¬ and Cosine Similarity methods. The output results in the form of 5 reviewer recommended names that can be selected by the software admin to be paired with the proposal.

(ii) This research succeeded in implementing TF-IDF and Cosine Similarity into the system for determining reviewers for research proposals for lecturers at the Indonesian University of Education.

(iii) The application of the TF-IDF and Cosine Similarity methods in recommending reviewers obtained 83.3% agreement based on the abstract and scientific field. From these results,

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088

this recommendation system can be said to be very good. Then in this research, it was found that when the scientific field of a proposer and reviewer is the same, but the abstract and the reviewer's scientific field are different, the reviewer's recommendations cannot be paired with the proposal.

5. REFERENCES

Badrul, M., and Utami, T. (2022). Penerapan metode profile matching untuk rekomendasi penunjang keputusan promosi jabatan Di PT. Inbisco Niagatama Semesta. *PROSISKO: Jurnal Pengembangan Riset dan Observasi Sistem Komputer, 9*(1), 14-20.

Darmawan, C. (2020). Implementasi kebijakan profesi guru menurut undang-undang republik Indonesia nomor 14 tahun 2005 tentang guru dan dosen dalam perspektif hukum pendidikan. *Wacana Paramarta: Jurnal Ilmu Hukum, 19*(2), 61-68.

Eskak, E. (2020, December). Kajian manfaat teknologi informasi dan komunikasi (tik) untuk meningkatkan daya saing industri kreatif kerajinan dan batik di era Industri 4.0. *In Prosiding Seminar Nasional Industri Kerajinan dan Batik, 2*(1), 10-10.

Indrayani, H. (2012). Penerapan teknologi informasi dalam peningkatan efektivitas, efisiensi dan produktivitas perusahaan. *Jurnal El-Riyasah, 3*(1), 48-56.

Kohila, R., and Arunesh, D. K. (2016). Text Mining: Text Similarity measure for news articles based on string-based approach. *Global Journal of Engineering Science and Research Management, 3*(7), 35-42.

Kurniawan, A., Solihin, F., and Hastarita, F. (2014). Perancangan dan Pembuatan Aplikasi Pencarian Informasi Beasiswa dengan Menggunakan Cosine Similarity. *Jurnal Simantec, 4*(2), 115-124.

Kurniawan, B., Effendi, S., and Sitompul, O. S. (2012). Klasifikasi konten berita dengan metode text mining. *Jurnal Dunia Teknologi Informasi, 1*(1), 14-19.

Mariel, W. C. F., Mariyah, S., and Pramana, S. (2018). Sentiment analysis: a comparison of deep learning neural network algorithm with SVM and naïve Bayes for Indonesian text. *In Journal of Physics: Conference Series, 971*(1), 012049.

Praniffa, A. C., Syahri, A., Sandes, F., Fariha, U., Giansyah, Q. A., and Hamzah, M. (2023). Pengujian sistem informasi parkir berbasis web pada UIN Suska Riau menggunakan white box dan black box testing. *Jurnal Testing dan Implementasi Sistem Informasi, 1*(1), 1-16.

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., and Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas, 18*(3), 491-504.

Siregar, R. R. A., Sinaga, F. A., and Arianto, R. (2017). Aplikasi Penentuan Dosen Penguji Skripsi Menggunakan Metode TF-IDF dan Vector Space Model. *Computatio: Journal of Computer Science and Information Systems, 1*(2), 171-186.

Sulistyo, M. E., Saptono, R., and Asshidiq, A. (2015). Penilaian ujian bertipe essay menggunakan metode text similarity. *Telematika: Jurnal Informatika dan Teknologi Informasi, 12*(2), 146-158.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2774-7891

Yasni, L., Subroto, I. M. I., and Haviana, S. F. C. (2018). Implementasi cosine similarity matching dalam penentuan dosen pembimbing tugas akhir. *Transmisi, 20*(1), 1-7.

Yuan, M., Ouyang, Y. X., and Xiong, Z. (2013). A text categorization method using extended vector space model by frequent term sets. *Journal of Information Science and Engineering, 29*(1). 99-114.

Zaware, S. N., Gautam, A., Nashte, S., and Khanuja, P. (2015). An effectual approach for calculating cosine similarity. *International Journal of Advance Engineering and Research Development, 2*(4), 13-18.

DOI: http://dx.doi.org/10.17509/xxxx.xxx
p- ISSN xxxx-xxxx e- ISSN 2723-4088