JEEMECS (Journal of Electrical Engineering, Mechatronic and Computer Science
Vol. 8, No. 1, February 2025, pp. 31-38
ISSN 2614-4859

Design and Build a Research Information System at National Cyber and Crypto Polythecnic with Recommender System for Thesis Supervisor Based on Text Similarity Metric

Rayhan Ramdhany Hanaputra [a,1], Hermawan Setiawan [b,2], Girinoto [b,3], Ray Novita Yasa [b,4], Rheva Anindya Wijayanti [b,5*], Jacob Lumbantoruan [b,6], Muhammad Lucky Aulia Firmansyah [b,7]

[a] Badan Siber dan Sandi Negara, Jalan Raya Muchtar, Depok 16518, Indonesia
[b] Politeknik Siber dan Sandi Negara, Jalan Raya H. Usa, Bogor 16120, Indonesia
[1] rayhan.ramdhany@bssn.go.id; [2] hermawan.setiawan@poltekssn.ac.id; [3] girinoto@poltekssn.ac.id;
[4] ray.novita@poltekssn.ac.id; [5] rhevaanindyaw@gmail.com*, [6] jacob.lumbantoruan@student.poltekssn.ac.id,
[7] muhammad.lucky@student.poltekssn.ac.id
* corresponding author

Keywords
Information system
Recommender system
Text similarity metric
Web scraping

ABSTRACT

Politeknik Siber dan Sandi Negara (Poltek SSN) currently does not have an adequate system for managing research proposals. As a university, it must fulfill the Tri Dharma of Higher Education, one of them is research. Therefore, it is necessary to build a proposal for a management information system. Several previous studies have shown that research information systems can be built with several web frameworks. The approach of utilizing web scraping technology can integrate Google Scholar research data into the information system and recommendation services for accompanying lecturers using the Text Similarity Metric method to provide relevant results. In this study, an integrated research information system application will be built with Google Scholar, implementing the Text Similarity Metric method for the recommendation system for accompanying lecturers, and using the CodeIgniter 4 framework. The integration of the system with this lecturer recommendation service will be tested through monitoring lecturer satisfaction related to the research services provided by the system developed. The results obtained in this study are a research system that works in accordance with the business processes at Poltek SSN with the highest accuracy score on the cosine similarity algorithm of 92,95% and user satisfaction test results of 97,76%.

1. Introduction

According to the Regulation of the State Cyber and Crypto Agency Number 12 of 2019, the State Cyber and Crypto Polytechnic (Poltek SSN) is a vocational higher education institution focusing on cyber security and cryptography under the State Cyber and Crypto Agency. As part of fulfilling the Tri Dharma of Higher Education, Poltek SSN emphasizes research as a core activity. Through its Center for Research and Community Service (PPM), Poltek SSN encourages lecturers and students to conduct and develop research in cyber security and coding to contribute to national security [1]. However, Poltek SSN currently lacks an information system to manage research proposals efficiently, necessitating the development of a research information system to streamline proposal submissions and provide a comprehensive repository of research activities [2].

A significant issue in the proposal submission process is selecting research assistant lecturers, often determined subjectively or based on departmental policies without aligning lecturers'

expertise with the research field. Since lecturers' competencies can be reflected in their publication track records on platforms like Google Scholar, integrating these records is crucial for improving this process. However, collecting and organizing this publication data is time-consuming [3]. Web scraping techniques for Google Scholar have gained attention for automating this process, supporting institutions like Poltek SSN in tracking publication performance and facilitating national directories like SINTA (Science and Technology Index) [4].

This research aims to design a research information system with a recommendation service for research assistant lecturers using Text Similarity Metrics based on Google Scholar web scraping results [5] [6]. Text similarity algorithms such as cosine similarity, Jaccard index, and others will be tested and compared for their performance using accuracy scores [7][8][9][10]. The best-performing metric will be implemented in the system, built using the CodeIgniter 4 framework. The system will optimize research proposal submission, document existing research, and assist researchers in finding suitable assistant lecturers [11]. Its effectiveness will be evaluated through satisfaction surveys with Poltek SSN lecturers and students.

2. The Proposed Method

Research information systems are the implementation of information technology to facilitate the research process, increase operational efficiency, and improve interaction and communication. This system records, stores, and processes data to support planning, operation, and supervision [12]. Research aims to obtain new information through discovery, proof, and development. Discovery ensures information is new, proof removes doubts, and development deepens existing knowledge [13][14][15]. The business process of the flow of submitting research proposals at Poltek SSN is depicted in Figure 1.

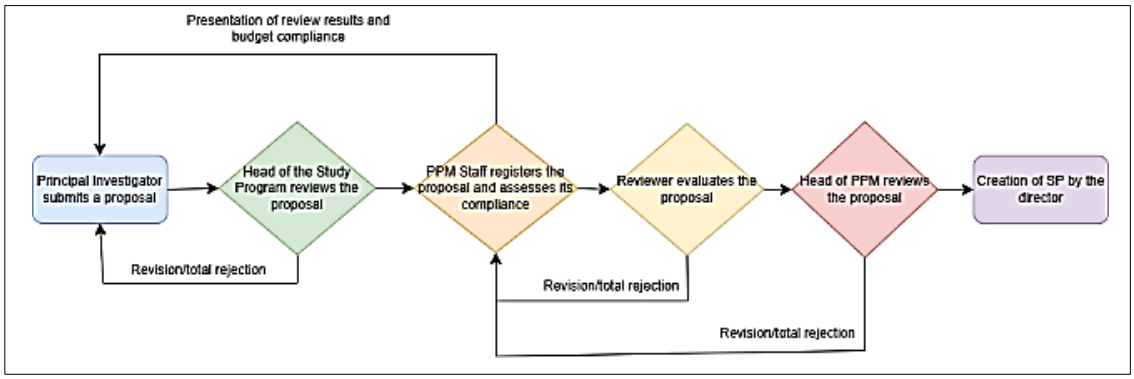

Fig. 1. Poltek SSN Research Proposal Flow Business Process

3. Method

The object of this research is the Poltek SSN Research and Community Service Center (PPM) with objects in the form of a Google Scholar integrated research information system application and a text similarity metric-based assistant lecturer recommendation service. The qualitative approach is used for information system development and web scraping, while the quantitative approach is applied to evaluate text similarity metric through accuracy score, where the best metric will be implemented in the system. The research methodology includes six stages: system design, Google Scholar web scraping, data preprocessing, model training, best model integration, and user satisfaction survey. The assessment of the results is done through a user satisfaction survey. The research flow is summarized in Figure 2.

 ISSN 2614-4859

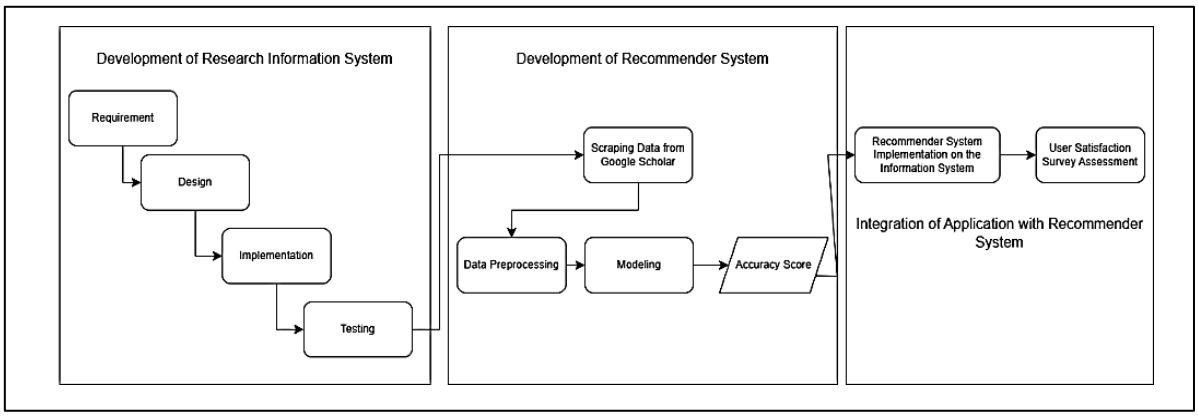

Fig. 2. Research Process Flow

At this stage, a research information system application will be built which also has a Google Scholar web scraping feature using the Simple HTML DOM Parser library. This system will facilitate the integration of research data with Google Scholar and also shorten the process of submitting proposals in the Poltek SSN environment. The application will be built using the waterfall model which consists of the stages of requirements, design, implementation, and testing. This model is suitable for projects with clear requirements and a fixed budget and schedule. It is also useful if the technology used is well understood and the researcher has experience with similar projects.

This stage identifies the problems and needs of the research information system, including features, functional, and non-functional requirements. The system is designed with Role-Based Access Control (RBAC) for four roles: user, admin, editor, and reviewer in Figure 3.

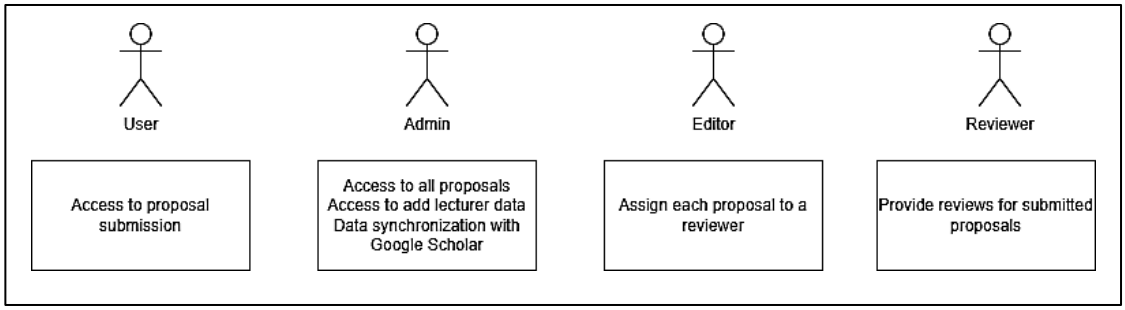

Fig. 3. Access roles on the information system

Then, the design stage will design the system workflow, database design, and Google Scholar data integration using UML (Use Case Diagram, Activity Diagram, Sequence Diagram) and ERD to describe data relations. Then, the implementation stage focuses on system development using the CodeIgniter 4 framework, MySQL database, and XAMPP v.3.3.0 on Windows 11, including the Google Scholar web scraping feature. Then, testing is done with User Acceptance Testing (UAT) to ensure the system meets business needs based on the needs that have been determined in the use case diagram as the testing stage.

The process involves designing the system workflow, database structure, and integrating Google Scholar data using UML (Use Case Diagram, Activity Diagram, Sequence Diagram) and ERD to illustrate data relationships. System development is implemented using the CodeIgniter 4 framework, a MySQL database, and XAMPP v.3.3.0 on Windows 11, incorporating a Google Scholar web scraping feature. Testing is conducted through User Acceptance Testing (UAT) to ensure the system aligns with business needs as outlined in the use case diagram.

Poltek SSN lecturer data is retrieved through Google Scholar scraping and stored in the research information system database for use in the training and prediction process of assistant lecturer recommendations. The scraped dataset is combined with student final project data from the Poltek SSN Library, followed by data preprocessing steps such as train-test data splitting, data cleaning, text conversion to lowercase, and stopwords removal to improve model accuracy. The text similarity model is built using Python with Jaccard Index, Euclidean Distance, and Cosine Similarity algorithms, and the model is evaluated based on accuracy, with the best-performing metric implemented in the system. This best metric is integrated into the system using Flask as a standalone API, allowing the model to accept research topic inputs and provide assistant lecturer recommendations to users. Finally, the system's effectiveness is assessed through a user satisfaction survey using a Likert scale (1-5) conducted with Poltek SSN lecturers and students, with the survey results used to evaluate how well the system addresses the problem formulation.

4. Results and Discussion

4.1. Development of Research Information System

This research produced a web-based research information system that was built using the Software Development Life Cycle (SDLC) methodology of the Waterfall model. This system was developed through the stages of Requirement, Design, Implementation, and Testing to support the research management process, synchronization of lecturer data, and recommendations for research assistant lecturers. The research information system developed aims to help manage research data by considering the needs of all users. The requirement stage defines functional and non-functional standards for the application, mapped into features listed in Table 1.

Table 1. Research Information System Feature Design

Number	*Feature*	*Function*
1.	Authentication Page	As a user login and registration page to authenticate to access research services.
2.	Dashboard Page	As a page accessed by users to see a brief summary of the data in the information system.
3.	Research Submission Page	As a page used by users to add, update, delete, and view research proposal submission data in the application.
4.	Document Scholar Page	As a page to view data on research documents conducted by Poltek SSN lecturers and integrated by Google Scholar.
5.	Lecturer Profile Page	As a page for managing Poltek SSN lecturer data that can be used by administrators in managing both adding, updating, deleting data, and synchronizing data with Google Scholar.
6.	Manage Account Page	As a user management center page used by administrators in adding, viewing, updating, and deleting user data on the information system.
7.	Profile Page	As a user profile management page used to view account information and change passwords.

The design stage uses UML (Use Case Diagrams, Scenarios, Activity Diagrams, Sequence Diagrams) and ERD to map entity relationships based on identified requirements. The implementation stage uses the CodeIgniter 4 framework, PHP, and MySQL, with a dashboard displaying research summaries like submitted proposals, citations, and most-cited lecturers on Google Scholar. The Dashboard page as in Figure 4 displays summary data, including the number of research proposals submitted, the number of citations per study program, as well as the most cited lecturers on Google Scholar.

System pages are tailored to user roles, enabling admins to manage proposals and lecturer profiles synced with Google Scholar, reviewers to assess assigned proposals, and users to track progress. UAT ensures all features meet user needs.

 ISSN 2614-4859

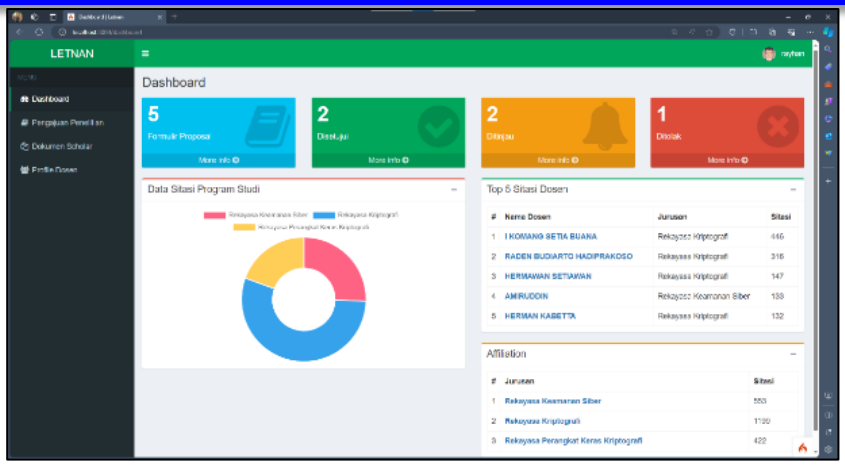

Fig. 4. Dashboard Page Display

Google Scholar data scraping involves manually inputting Poltek SSN lecturer data and Google Scholar IDs, with synchronization results stored in the database for further use. The preprocessing stage combines Google Scholar scraping results with final project research datasets, performing data cleaning and preparation. The modeling process then trains text similarity algorithms—Jaccard Index, Euclidean Distance, and Cosine Similarity—using a 90:10 train-test data ratio, with performance compared based on metrics such as in Table 2.

Table 2. Comparison of Evaluation Results

Evaluation Metric	Value		
	Jaccard Index	***Euclidean Distance***	***Cosine Similarity***
Accuracy	77.46%	42.25%	92.96%
Precision	65.48%	49.84%	84.03%
Recall	62.15%	35.94%	81.90%
F1-Score	61.36%	36.99%	81.79%

The results of modeling will get accuracy, precision, recall, and F1-score to compare the results of each metric. The data ratio that will be used in the training and test process is 9: 1, with 90% as train data and 10% as test data.

The best model is implemented in the research information system with a recommender system feature for research assistant lecturers, built using Python and Flask as an API to deliver recommendations via the application interface. The interface of the recommender system feature for research assistant lecturers can be seen in Figure 5

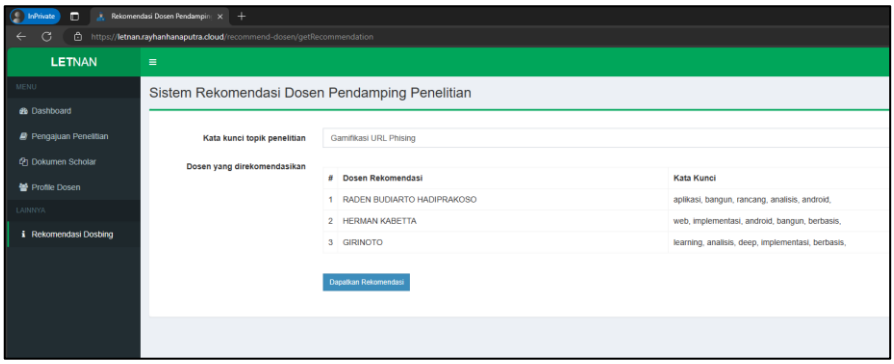

Fig. 5. Dashboard Page Display

A user satisfaction survey involving 83 lecturers and students showed a satisfaction level of 97.76%, with 4057 out of 4150 points accumulated after testing the application.

5. Conclusion

The results showed that the designed Poltek SSN research information system has functioned in accordance with the business process at the Poltek SSN Research and Community Service Center (PPM), with features that facilitate the submission, review, and monitoring of research proposals. The cosine similarity algorithm, with an accuracy of 92.95%, proved to be the best metric for the recommender system in recommending assistant lecturers, based on a comparison with the Jaccard index and Euclidean distance using the Poltek SSN research dataset. The comparison of datasets used as train data and test data is 9:1 from the Poltek SSN research dataset obtained through Google Scholar and also the Poltek SSN Library Unit. This information system also succeeded in meeting user expectations, with a satisfaction level of 97.76% from 83 respondents of Poltek SSN lecturers and cadets. Some suggestions for improving the information system application include improving the appearance to make it more attractive and adding a proposal process tracking feature based on the results of user satisfaction surveys, and developing a recommender system using additional data other than research titles and applying machine learning algorithms.

References

[1] I. Chudzaifah, A. N. Hikmah, and A. Pramudiani, "Tridharma Perguruan Tinggi: Sinergitas Akademisi dan Masyarakat dalam Membangun Peradaban," *Al-Khidmah J. Pengabdi. dan Pendampingan Masy.*, vol. 1, no. 1, pp. 79–93, 2021.

[2] R. Agustino, Y. B. Widodo, A. Wiyatno, and M. I. Saputro, "Sistem Informasi Penelitian dan Pengabdian Masyarakat di Universitas Mohammad Husni Thamrin: Sistem Informasi LPPM, Sistem Database LPPM, Rancang Bangun LPPM," *J. Jaring SainTek*, vol. 2, no. 1, pp. 1–12, 2020.

[3] M. Asfi and N. Fitrianingsih, "Implementasi Algoritma Naive Bayes Classifier sebagai Sistem Rekomendasi Pembimbing Skripsi," *InfoTekJar J. Nas. Inform. dan Teknol. Jar*, vol. 5, no. 1, pp. 44–50, 2020.

[4] A. Rahmatulloh and R. Gunawan, "Web scraping with HTML DOM method for data collection of scientific articles from Google Scholar," *Indones. J. Inf. Syst.*, vol. 2, no. 2, pp. 95–104, 2020.

[5] A. F. Putri, G. Manik, F. Nabila, and N. Chamidah, "Implementasi Scraping Google Scholar Menggunakan HTML DOM Untuk Pengumpulan Data Artikel Dosen UPN Veteran Jakarta Berbasis Web," in *Prosiding Seminar Nasional Mahasiswa Bidang Ilmu Komputer Dan Aplikasinya*, 2021, vol. 2, no. 1, pp. 668–678.

[6] H. Steck, L. Baltrunas, E. Elahi, D. Liang, Y. Raimond, and J. Basilico, "Deep learning for recommender systems: A Netflix case study," *AI Mag.*, vol. 42, no. 3, pp. 7–18, 2021.

[7] D. D. Prasetya, A. P. Wibawa, and T. Hirashima, "The performance of text similarity algorithms," *Int. J. Adv. Intell. Informatics*, vol. 4, no. 1, pp. 63–69, 2018.

[8] A. Azhari, E. Buulolo, and N. Sialalhi, "Sistem Rekomendasi Dosen Pendamping Skripsi Berbasis Text Rank menggunakan Metode Cosine Similarity," *Pelita Inform. Inf. dan Inform.*, vol. 10, no. 3, pp. 119–122, 2022.

[9] A. W. Qurashi, V. Holmes, and A. P. Johnson, "Document processing: Methods for semantic text similarity analysis," in *2020 international conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 2020, pp. 1–6.

[10] R. Singh and S. Singh, "Text similarity measures in news articles by vector space model using NLP," *J. Inst. Eng. Ser. B*, vol. 102, pp. 329–338, 2021.

[11] M. Alfano, A. E. Fard, J. A. Carter, P. Clutton, and C. Klein, "Technologically scaffolded atypical cognition: The case of YouTube's recommender system," *Synthese*, vol. 199, pp. 835–858, 2021.

[12] E. P. Primawanti and H. Ali, "Pengaruh Teknologi Informasi, Sistem Informasi Berbasis Web Dan Knowledge Management Terhadap Kinerja Karyawan (Literature Review Executive Support Sistem (Ess) for Business)," *J. Ekon. Manaj. Sist. Inf.*, vol. 3, no. 3, pp. 267–285, 2022.

[13] N. F. Fauziah and E. Retnoningsih, "Sistem Informasi Penelitian dan Pengabdian kepada Masyarakat Berbasis Web," *Inf. Manag. Educ. Prof. J. Inf. Manag.*, vol. 4, no. 2, pp. 183–192, 2020.

[14] A. Musrifah, "Rancangan Sistem Informasi Penelitian Dan Pengabdian Kepada Masyarakat Di Fakultas Teknik Universitas Suyakancana," *INFOTECH J.*, vol. 8, no. 1, pp. 5–12, 2022.

[15] M. Ridwan, A. M. Suhar, B. Ulum, and F. Muhammad, "Pentingnya penerapan literature review pada penelitian ilmiah," *J. Masohi*, vol. 2, no. 1, pp. 42–51, 2021.

ISSN 2614-4859

This Page Is Intentionally Left Blank

http://jurnal.unmer.ac.id/index.php/jeemecs jeemecs@unmer.ac.id