Information Retrieval System for Scientific Publications of Lampung University by using VSM, K-Means, and LSA

Rahman Taufik, Didik Kurniawan*, Anie Rose Irawati, Dewi Asiah Shofiana
Dept. of Computer Science, University of Lampung, Bandar Lampung, Indonesia

***Abstract*—The Lampung University repository system is a repository of data related to study, community service, and other scientific works, currently has 37242 documents accessible through repository.lppm.unila.ac.id. Despite the amount of data, its optimal use as an information retrieval remains unrealized, hindering the effective promotion of Lampung University's scientific publication excellence. Recognizing the limitations of existing information retrieval systems that are limited to specific methods for topic identification through clustering, this study aims to develop a retrieval system for Lampung University's repository using Vector Space Model (VSM), K-Means and Latent Semantic Analysis (LSA) that generates clusters and study expertise at the level of study program, faculty and Lampung University. The methodology includes data collection, preprocessing, modeling, evaluation and system deployment. The results show that the number of clusters obtained for the university level is 7 clusters, for the faculty level are 6, 7, 8 and 10 clusters, and for the program level are 3 to 5 clusters. In addition, the finding topic identification indicate that the expertise topics at Lampung University, which are agriculture, soil, education, plants, learning, society, Lampung. This study contributes to optimizing the information retrieval system, promoting academic excellence, and advancing the understanding of study expertise at Lampung University.**

Keywords*—Information retrieval; Vector Space Model (VSM); k-means; Latent Semantic Analysis (LSA); clustering; topic identification; scientific publication information

I. INTRODUCTION

The Lampung University repository system is a data repository accessible to all lecturers and the academic community through the address repository.lppm.unila.ac.id. According to Google index data, the publication data already contains 37242 documents. Additionally, the University of Lampung Repository System is utilized to store lecturer data related to study activities, publications, community service, and other achievements. All data stored in this repository is traceable based on divisions at the University of Lampung, namely faculties, study programs, authors and years of study. The data stored in this repository can serve various purposes, including as a source of information, for literature reviews, accreditation, self-evaluation, institutional achievements, and other *tri dharma*-related purposes.

However, the data in the repository has not been effectively utilized as meaningful information due to the inadequacy of the system in accessing and managing data in an informative manner. Consequently, units within Lampung University encounter difficulties when searching for relevant scientific publications in their areas of expertise. Furthermore, understanding the strength of study fields based on the obtained information is crucial to assess the excellence of study at the University of Lampung.

One of the technologies that can be used to extract information from a repository system is an information retrieval system. Several studies such as Information Retrieval for Digital Library [1], Information Retrieval for Faculty Study Repositories [2], Information Retrieval for Bibliographic Control and Institutional Repositories [3], Multilingual Information Retrieval [4], have proven that information retrieval systems can effectively collect and determine information from various repositories. In study [1], they improved the existing digital library system, namely Sowiport, by integrating heterogeneous databases through an information retrieval system approach. As a result, they obtained 513,000 data entries from 25 different thesauri, enhancing keyword search capabilities. In addition to digital library system [1], there are other studies related to educational repositories which are studies [2][3]. The study [2] aims to evaluate faculty study repositories used in higher education institutions. One aspect of the evaluation is to identify relevant articles from nine academic databases using an information retrieval system approach. The results indicate that the evaluated and redesigned system improves the preservation of scientific study results. While study [3] improves the quality of their data, metadata, and semantic data through an information retrieval system approach, the results support their control over the bibliography, including a repository that is rich and clean. In terms of diversity of information data, compared to other studies that collect data in various languages through information retrieval systems [4], they successfully gathered over 77 thousand Wikipedia queries in 18 languages. This supports improvements in data retrieval across multiple languages.

*Corresponding Author

The implementation of an information retrieval system must be accompanied by the utilization of techniques. The follow-up regarding the application of techniques can be observed in studies [5] [6] [7]. Study [5] employs data mining techniques to enhance information retrieval within their digital institutional repository, yielding personalized profiles specific to each user group in their repository. In study [6], an information retrieval system and data mining techniques were employed, specifically an Expertise Recommender System. This approach was utilized to develop a prototype system aimed at identifying and recommending thesis advisors for specific subjects. Unfortunately, specific data mining methods used in studies [5] [6] are not mentioned. On the other hand, study [7] developed an information retrieval system using hybrid deep fuzzy hashing algorithm to address issues related to similarity measurement procedures. Experimental result showed that the proposed model achieved higher retrieval accuracy compared to conventional models, although it is limited in feature selection.

Additionally, other approaches, such as Natural Language Processing (NLP), can be employed, as demonstrated in study [8] [9]. Study [8] developed an information retrieval system using NLP techniques to address employee queries based on knowledge from their internal site. The results showed that the queries were successfully responded to with relevant answers, although there is potential for further acceleration in the process. Furthermore, study [9] developed a search system by implementing information retrieval system techniques, Jaccard and cosine similarity matrix techniques. These methods were utilized to clean and standardize data and information related to academic documents. The results indicated that the system could be utilized to recommend documents, albeit with limitations in searches based on divergence metrics.

Despite significant advancements in the utilization of information retrieval systems across various domains [1] [2] [3] [4], there appears to be a noticeable gap in explicit discussions or study concerning the optimization of data generated by information retrieval systems. Several studies [5] [6] [7] [8] [9] mention the use of data mining techniques and NLP, but specific methods for topic identification through clustering have not been proposed. The implementation of clustering for topic identification is crucial, particularly in the context of academic data.

The present study aims to develop an information retrieval system for scientific publications of Lampung University using the Vector Space Model (VSM), K-Means and Latent Semantic Analysis (LSA). We propose VSM to obtain the similarity in vector representations, K-Means to cluster study topics, while LSA is used to generate the identification of study topics. The use of Vector Space Model (VSM) method can improve the information obtained from information retrieval systems. A number of studies have developed information retrieval system that applied VSM to determine information [10] [11]. The study in [10] proposed the combination of two methods, namely VSM and description logic for concept extraction, which increases the level of similarity between documents and certain concepts in information retrieval systems. While, the study [11] proposed VSM based on TF-IDF weights and word vectors to calculate semantic similarity for implicit citation detection problem. On the other hand, studies such as [12] [13] show that K-means method can be proposed for clustering scientific publication documents. The study [12] proposes a multi-verse optimizer and k-means algorithm for extracting topics from clustered documents. K-means can be used especially in cases where the dataset is normally distributed such as scientific publication documents [13]. Furthermore, regarding LSA, study [14] proposed a method for information retrieval systems using LSA method to retrieve important information from questions asked by users or mass documents, and the results showed a positive contribution. Therefore, the selection of VSM, K-means, and LSA is proposed based on a literature review [10] [11] [12] [13] [14], given their effectiveness in addressing the specific aspects required for the study, the use of alternative methods may not align with our objectives.

The objective of this study is to develop an information retrieval system that generates clusters and study expertise at the level of study program, faculty and Lampung University. The study is driven by the following three research questions: (1) How to develop the information retrieval system for scientific publications using the Vector Space Model (VSM), K-Means and Latent Semantic Analysis (LSA)? (2) How many clusters are formed based on data generated by the information retrieval system? (3) What is the most dominant study topics based on the results of the information retrieval system? Answers to these study questions are provided and discussed in the remaining sections of this paper, particularly in the results and discussion section. We aim to contribute to the ongoing discourse on information retrieval system development and provide practical insights on VSM, K-Means and LSA implementation to determine study expertise.

II. STUDY METHODOLOGY

In developing the information retrieval system (see Fig. 1), various stages are proposed, which include data collection, preprocessing, modeling, evaluation, and system distribution. Initially, 37242 scientific publication documents from lecturers were collected from the repository of the Institute for Study and Community Service (LPPM), University of Lampung. After collection, the data was preprocessed by keyword extraction, cleaning, language detection, translation, stemming, and stopword removal. The data was then modeled to achieve clustering and classification of study topics. The evaluation was done using clustering test methods such as Silhouette score [15], Calinski-Harabasz score [16], and Davies-Bouldin score [17]. Finally, the obtained study data on clusters and scientific publication expertise were distributed as meaningful information for Lampung University.

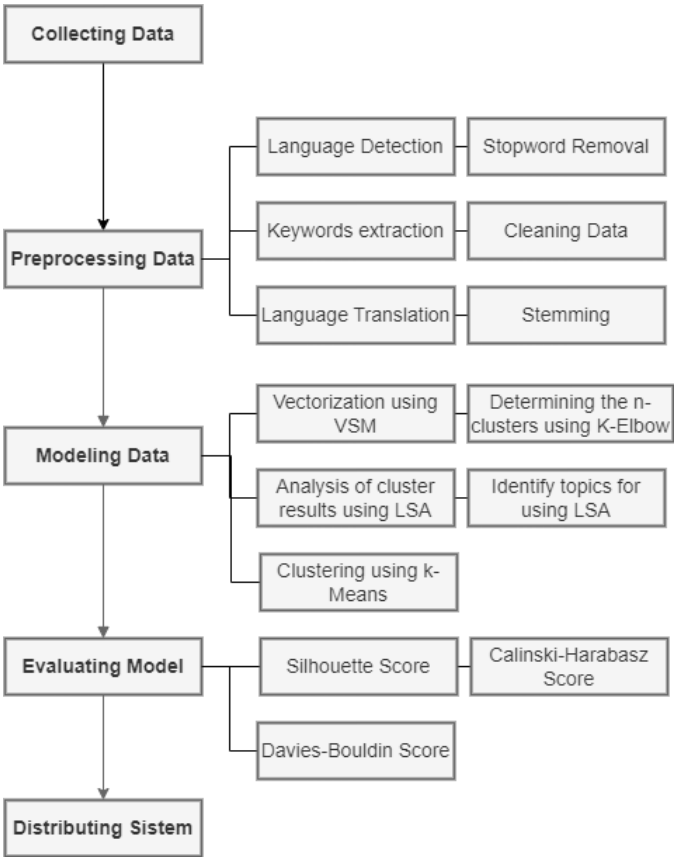

Fig. 1. Stages of information retrieval system development.

III. Results and Discussion

A. Data Collection

The first stage in the development of the Lampung University Lecturer Publication Information Retrieval System is literature study and data collection. Data collection is done by first creating a program using the Python programming language which can be seen in Fig. 2. This program was created to retrieve raw data from the Lampung University's LPPM repository (http://repository.lppm.unila.ac.id/) into collection data. The rest, this program is used to collect data from each URL that has been registered, clean the data collection and convert it into csv.

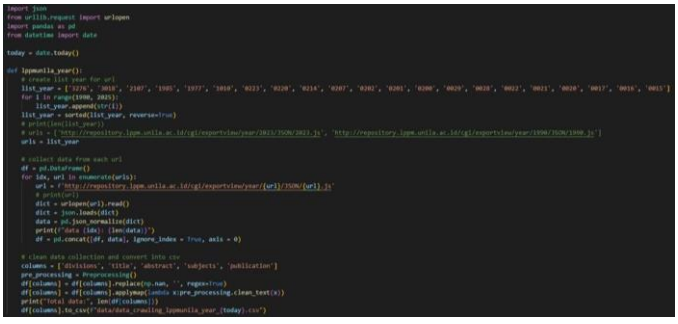

Fig. 2. Code of the study publication data collection program.

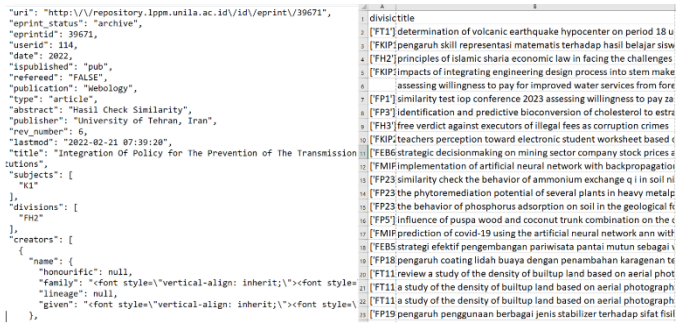

Fig. 3. Data collection of study publication.

Using the above program, 37242 data were retrieved. These data are scientific publication data of Lampung University lecturers, including division, title, abstract, subjects, publication, publisher, date, keywords. Fig. 3 shows raw data from Lampung University's LPPM repository and collection data in the form of .csv files processed by the program.

B. Data Preprocessing

The next step is data preprocessing. This stage aims to clean the collection data prepared for data modeling. This stage includes several activities, including keyword extraction, data cleaning, language detection, language translation, stemming, and stopword removal. These preprocessing stages are performed sequentially.

The collection data obtained from the Lampung University's LPPM repository does not have good consistency, some data have keywords and some do not, so the first preprocessing stage performed is the separation of keywords from abstract data using the Python library and regex. After the keywords data are obtained, the title and abstract data are cleaned by separating some unimportant characters and symbols. In addition to inconsistencies related to keywords, the collection data obtained is also inconsistent regarding language, therefore the next stage is to detect the language, and if there is data in English then it is translated into Indonesian. After all the data is in Indonesian, the next data is stemming which converts a word to its root word by removing the phrase prefix [18]. The last preprocessing stage performed is stopword removal; this stage is performed to remove unimportant words. The stopword removal can be seen in Fig. 4. Furthermore, the preprocessing results of keywords, abstract and title data are combined into a feature that is ready to be used for data modeling.

Fig. 4. The stopword removal to remove unimportant words..

C. Data Modeling

The next stage is data modeling which is the core of the development of this information retrieval system. Data modeling in this study is aimed at obtaining clustering and topic classification from study data using various machine learning methods. The modeling stages include text vectorization using the VSM model, determining the number of clusters using the k-elbow method, clustering using the k-means method, visual analysis of cluster results using the Umap library and the LSA method, and topic classification using the LSA method. This data modeling stage is performed on three different types of data, first for university level data, second for faculty data and third for study program data.

The collection data are preprocessed and then converted into vector form using VSM. One of the VSM methods used is TF-IDF (Term Frequency - Inverse Document Frequency). TF-IDF produces a vector value based on measuring the authenticity of a word by comparing the number of occurrences of a word in a document with the number of occurrences of a document containing that word. The vector results obtained from the University of Lampung study data have vector dimensions (37242, 3094428). The result of the vector is then used to determine the number of clusters and the clustering process, k-elbow and k-means methods are proposed in this case. In the results, the number of clusters obtained for the university level is 7 clusters, while for the faculty level is 6, 7, 8 and 10 clusters, while for the study program level is 3 to 5 clusters. The details for the faculty and study program levels can be seen in Tables I and II. In addition, the results of clustering at the university level can be seen visually in Fig. 5.

TABLE I. FACULTY-LEVEL CLUSTERING INFORMATION

Faculty	Total Data	Cluster Number	Topic
FKIP (Faculty of Teacher Training and Education)	4710	8	student, learning, teacher, education, teaching, Lampung, language,class
FMIPA (Faculty of Mathematics and Natural Sciences)	5042	7	plant, orchid, plant, extract, method, acid, virus
FEB (Faculty of Economics and Business)	2207	8	business, work, Lampung, indonesia, employee, consumer, tax,money
FT (Faculty of Engineering)	4401	6	land, water, city, oil, material, earth
FP (Faculty of Agriculture)	9022	10	plant, forest, agriculture, fertilizer, tree, food, water, plant, food,Lampung
FISIP (faculty of Social Sciences and Political Sciences)	2912	7	lampung, village, community, travel, tourism, group, work, district
FK (Faculty of Medical)	6569	8	patient, health, mind, development, behavior, school
FH (Faculty of Law)	2466	7	law, custom, community, protected, Lampung, tax, data

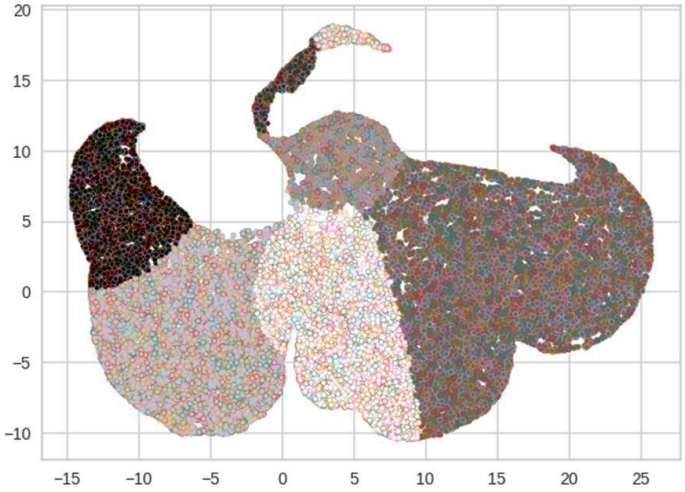

Fig. 5. Visualization of University-level clustering.

In addition to clustering, the results of this study also generate a set of topic identification from each clustering obtained. These topics identifications are generated using the LSA method. For each topic identification from the clusters obtained, then a topic is selected that is very relevant in the context of the topic in manually, for example from the Faculty of Agriculture 2 clusters are produced, the first cluster has classified topics including study, results, farming, while the second cluster is soil, processing, farming, then from the first cluster the topic taken is farming and the second cluster the topic taken is soil. This is because the topic represents the cluster; in addition to the words farming and soil have a context related to the study topics of the Faculty of Agriculture. The results of topic classification for the university level include agriculture, soil, education, plants, learning, society, Lampung. As for faculties and study programs can be seen in Table I and Table II.

D. Data Evaluation

The next stage is model evaluation. The model evaluation stage is used to test the quality of the clustering results of the proposed model. Several methods are used to obtain the evaluation results, including Silhouette score, Calinski-Harabasz score, and Davies Bouldin score. The Silhouette score with a value close to 1 means that the data point is in the correct cluster, close to -1 means that the data points is in wrong cluster [15]. While in Calinski-Harabasz calculations, the higher the value, the better the grouping is [16]. On the other hand, for the Davies Bouldin score, the smaller the value produced, the more optimal the clustering model [17]. The clustering model evaluation results obtained for the university level are 0.53 for the Silhouette score, 107298.36 for the Calinski Harabasz score, and 0.60 for the Davies Boulding score. The plot of these scores shows that the cluster model formed is quite good. Meanwhile, the results of the clustering model evaluation for the faculty and study program levels can be seen in Table III and Table IV.

TABLE II. STUDY PROGRAM-LEVEL CLUSTERING INFORMATION

Faculty	ID Study Program	Study Program	Total Data	Cluster Number	Topic
FEB	FEB1	Master of Accounting ScienceProgram	17	5	managerial, organization, service, government, audit
FEB	FEB2	Master of Economics Program	35	3	economy, employment,decentralization
FEB	FEB3	Master of Management Program	72	3	investment, loyalty, funding
FEB	FEB4	Accounting Study Program	731	4	finance, work, management,tax, financial
FEB	FEB5	Development Economics StudyProgram	465	5	economy, umkm, labor,investment, finance
FEB	FEB6	Management Study Program	899	4	ownership, consumer, production, business, work
FH	FH1	Doctoral Program in Law	880	3	law, politics, authority
FH	FH2	Master of Law Program	316	4	property rights, law, certificate,law
FH	FH3	Law Study Program	1271	4	customary law, rights, tax,village
FISIP	FISIP1	Business Administration Program	319	5	business, quality, banking,management, work
FISIP	FISIP2	State Administration Program	980	5	tourism, development, village,policy, program
FISIP	FISIP3	Government Science Program	621	4	government, politics,corruption, facilities
FISIP	FISIP4	Communication Study Program	379	5	communication, information, regression, behavior, media
FISIP	FISIP5	Master of Administrative ScienceProgram	13	4	management, work, fluctuation,policy
FISIP	FISIP6	Master of Government ScienceProgram	125	5	organization, culture, politics,conflict, strategy
FISIP	FISIP7	Sociology Study Program	356	4	society, conflict, behavior,crime
FISIP	FISIP8	International Relations Program	124	4	cooperation, country, export,relationship
FK	FK1	Medical Education Study Program	1860	4	patient, disease, treatment,doctor
FKIP	FKIP1	English Language Program	2202	4	students, classes, teachingtechniques, learning
FKIP	FKIP10	Master's Program in SocialStudies Education	96	4	development, teaching, social,student
FKIP	FKIP11	Master's Program in EducationalTechnology	177	4	students, learning model,motivation, education
FKIP	FKIP12	Civics Study Program	161	4	character, nation, education,culture
FKIP	FKIP14	Indonesian and Regional Language and Literature Education Study Program	174	5	language, speech, society,teaching, indonesia
FKIP	FKIP15	Biology Education Program	271	4	students, skills, science,learning model
FKIP	FKIP16	Social Studies EconomicsEducation Program	159	3	students, economy,entrepreneurship
FKIP	FKIP17	Physics Education Program	386	5	students, physics, skills, development, learning model
FKIP	FKIP18	Social Studies GeographyEducation Program	191	4	student, soil, factor, landslide
FKIP	FKIP19	Elementary School Teacher Education Study Program (PGSD)	143	4	students, education, development, teaching
FKIP	FKIP2	Guidance and Counseling StudyProgram	1386	4	students, learning model,concentration, education
FKIP	FKIP20	Physical Education, Health andRecreation	54	5	student, sport, ball, athlete,learning model
FKIP	FKIP21	Chemistry Education Program Management	476	4	students, experiment, teaching,learning model
FKIP	FKIP22	Mathematics Education Program	265	4	students, math, diagram,learning model
FKIP	FKIP23	PG-PAUD Education Program	183	4	child, development, social,behavior
FKIP	FKIP24	Social Studies History EducationProgram	159	3	history, students, education
FKIP	FKIP25	Drama, Dance and MusicEducation Program	150	5	music, dance, art, performance,works
FKIP	FKIP3	Master's Program in ElementaryTeacher Education	111	4	empirical, student, development, teaching
FKIP	FKIP4	Master of Science TeacherTraining Program	169	4	students, systems, learningmodels, skills
FKIP	FKIP5	Master's Program in Education	150	4	school, leadership, management, student
FKIP	FKIP6	Master's Program in EducationManagement	20	3	school, teacher, education
FKIP	FKIP7	Master's Program in Regional Language and Literature Education	62	4	language, semantics, tradition,teaching
FKIP	FKIP8	Master's Program in Indonesian Language and Literature Education	103	4	teaching, language, students,literature
FKIP	FKIP9	Master's Program in PhysicsEducation	508	4	students, learning models,development, skills
FMIPA	FMIPA1	Physics Study Program	415	4	temperature, data, method,concentration
FMIPA	FMIPA10	Information Systems Program	18	4	system, application, information, classification

Faculty	ID Study Program	Study Program	Total Data	Cluster Number	Topic
FMIPA	FMIPA2	Biology Study Program	1724	4	plant, extract, plant, orchid
FMIPA	FMIPA3	Computer Science Program	518	5	system, information, medicine,recommendation
FMIPA	FMIPA4	Chemistry Study Program	1079	4	compound, ion, village,adsorption
FMIPA	FMIPA5	Master's Program in BiologicalSciences	127	4	extract, steroid, taurine, benzo
FMIPA	FMIPA6	Master's Program in PhysicalScience	33	3	sample, temperature, point
FMIPA	FMIPA7	Master's Program in ChemicalScience	75	3	mushroom, silica, synthesis
FMIPA	FMIPA8	Master's Program in MathematicalSciences	46	3	model, probability, test
FMIPA	FMIPA9	Mathematics Study Program	1033	4	method, model, data, students
FP	FP1	Agribusiness Study Program	3758	5	food, community, agriculture,water, forest
FP	FP11	Master of Forestry ScienceProgram	217	4	forest, management, community, diversity
FP	FP12	Master of Environmental ScienceProgram	72	5	health, environment, social,community, ecotourism
FP	FP13	Master's Program in Development Counseling / Community Empowerment	18	3	sanitation, counseling,community
FP	FP14	Master's Program in Agricultural Extension and Communication	12	5	agriculture, village, extension,farmer, citizen
FP	FP15	Master's Program in AgriculturalIndustrial Technology	60	4	processing, coffee, flowers,technology analysis
FP	FP17	Animal Husbandry Study Program	389	4	livestock, livestock products, methods, animal diseases
FP	FP18	Agricultural Engineering Program	973	5	fertilization, moisture content, cropping techniques, agricultural yield, processing
FP	FP19	Agricultural Product TechnologyProgram	784	4	products, processing, agricultural products, processing
FP	FP2	Agro technology Study Program	2941	4	fertilization, agriculture, yieldprocessing, tillage
FP	FP20	Aquatic Resources Program	108	4	fish, water, fry, care
FP	FP21	Marine Science Program	95	4	sea, fish, aquaculture, care
FP	FP22	Agricultural IndustrialTechnology Program	136	3	agricultural products, processing, industrial innovation
FP	FP23	Soil Science Program	516	4	soil, fertilization, tillage,planting method
FP	FP24	Plant Protection Study Program	458	4	plant diseases, care, treatment,prevention
FP	FP25	Agronomy and Horticulture StudyProgram	614	4	crops, fertilization, seeds, pests
FP	FP26	Agricultural Extension Program	507	4	agriculture, extension, farmers,agricultural products
FP	FP27	Animal Nutrition and FeedTechnology Program	97	4	livestock products, nutrition,innovation, processing
FP	FP3	Aquaculture Study Program	375	4	aquaculture, fish, shrimp, care
FP	FP4	Doctoral Program in AgriculturalSciences	12	3	treatments, plant diseases, riskanalysis
FP	FP5	Forestry Study Program	1871	4	forest, community, forestproducts, fauna
FP	FP6	Master of Agribusiness Program	31	3	organization, result analysis,method analysis
FP	FP7	Master Program in Agroecotechnology	17	4	plants, replication, innovationresults, maintenance
FP	FP8	Master Program in Agronomy	32	4	crop, experiment, fertilizerprocessing, development
FT	FT1	Geophysical Engineering Program	1353	4	soil, rock, reservoir, water
FT	FT11	Geodetic Engineering StudyProgram	146	4	land, measurement, mapping,data collection
FT	FT12	D3 Civil Engineering StudyProgram	13	5	settlement, community, social, development, sustainability
FT	FT13	D3 Survey and Mapping StudyProgram	41	4	mapping, survey, community,evaluation
FT	FT14	D3 Mechanical Engineering StudyProgram	34	3	electrical, engine, testing
FT	FT2	Master Program in CivilEngineering	116	4	concrete, measurement, testing,material
FT	FT3	Master's Program in MechanicalEngineering	41	4	mechanism, machine,processing, material
FT	FT4	Civil Engineering Program	843	4	soil, rainfall, drainage, roads
FT	FT5	Electrical Engineering Program	528	4	electrical, voltage, network,sensor
FT	FT6	Chemical Engineering Program	553	3	waste, treatment, processing
FT	FT7	Mechanical Engineering Program	549	5	machinery, materials, temperature regulation, processing, measurement
FT	FT8	Informatics Engineering Program	231	4	information system, development, sensor, network
FT	FT9	Architecture Program	200	4	building, architecture, publicspace, tourism

TABLE III. FACULTY-LEVEL CLUSTERING EVALUATION RESULTS

Faculty	*Silhouette score*	*Calinski Harabasz score*	*Davies Bouldin Score*
FKIP (Faculty of Teacher Training and Education)	0.4109131575	8379.067221	0.7750398559
FMIPA (Faculty of Mathematics and Natural Sciences)	0.4897095986	19386.7335	0.6157010033
FEB (Faculty of Economics and Business)	0.4059269475	4079.690594	0.7504302209
FT (Faculty of Engineering)	0.4865720672	13639.76272	0.5176474339
FP (Faculty of Agriculture)	0.4546237817	13403.42462	0.7251143872
FISIP (faculty of Social Sciences and Political Sciences)	0.5142060504	8154.666587	0.5947357487
FK (Faculty of Medical)	0.4211412604	11129.5431	0.7477032403
FH (Faculty of Law)	0.476002308	7443.539066	0.7927361434

TABLE IV. STUDY PROGRAM-LEVEL CLUSTERING EVALUATION RESULTS

No	Study Program	*Silhouette score*	*Calinski Harabasz score*	*Davies Bouldin Score*
1	Master of Accounting ScienceProgram	0.44474428	26.25916184	0.5797495286
2	Master of Economics Program	0.9106475354	455.880385	0.07888671825
3	Master of Management Program	0.6558252049	195.8234061	0.4961123569
4	Accounting Study Program	0.4686553933	1168.559533	0.6627273443
5	Development Economics StudyProgram	0.5203163725	1013.045096	0.5052247982
6	Management Study Program	0.5370269291	2506.570266	0.5440162171
7	Doctoral Program in Law	0.712811619	3209.953671	0.3392638925
8	Master of Law Program	0.6382238705	716.6067985	0.4373224099
9	Law Study Program	0.4513430856	2772.691672	0.6360769562
10	Business Administration Program	0.4743366388	379.7126064	0.6451603922
11	State Administration Program	0.4971081968	2627.913425	0.6290085573
12	Government Science Program	0.5262378928	1065.168098	0.5217246815
13	Communication Study Program	0.5603440786	647.3738657	0.5011094346
14	Master of Administrative ScienceProgram	0.3540879556	40.03367626	0.7137427753
15	Master of Government ScienceProgram	0.5229085354	290.8637055	0.5319551749
16	Sociology Study Program	0.484876088	739.6636388	0.6474297144
17	International Relations Program	0.5733409428	498.2796745	0.4034342429
18	Medical Education Study Program	0.4992342555	2733.869864	0.5854277671
19	English Language Program	0.4825049699	3808.495565	0.6538096004
20	Master's Program in SocialStudies Education	0.4877305057	246.2186114	0.5998584388
21	Master's Program in EducationalTechnology	0.5661563429	360.5946676	0.5367123514
22	Civics Study Program	0.4759859002	256.7458577	0.5885390139
23	Indonesian and Regional Language and Literature Education Study Program	0.4261095962	292.1032256	0.7184361478
24	Biology Education Program	0.47853607	333.9577153	0.7308542766
25	Social Studies EconomicsEducation Program	0.5330673746	225.2167718	0.5502047002
26	Physics Education Program	0.4742562866	1062.980569	0.6636898044
27	Social Studies GeographyEducation Program	0.4923488311	448.1597782	0.6056313622
28	Elementary School Teacher Education Study Program (PGSD)	0.4217206704	194.523634	0.7536003486
29	Guidance and Counseling StudyProgram	0.5102595257	2887.425063	0.5507603067
30	Physical Education, Health andRecreation	0.6649952221	196.0591115	0.3759244643
31	Chemistry Education Program Management	0.4990151929	1161.093757	0.5890013265
32	Mathematics Education Program	0.5470512319	512.7211607	0.5589964858
33	PG-PAUD Education Program	0.5113909281	317.8699604	0.5416860298
34	Social Studies History EducationProgram	0.5890686005	479.1019235	0.4484316617
35	Drama, Dance and Music Education Program	0.3787565327	158.6606907	0.8101570708
36	Master's Program in Elementary Teacher Education	0.5726408051	244.2289309	0.4642264092
37	Master of Science TeacherTraining Program	0.5083243542	390.5031415	0.569395827
38	Master's Program in Education	0.526775473	410.7632211	0.5189634631
39	Master's Program in EducationManagement	0.7068854737	154.2750311	0.3192405471

No	Study Program	*Silhouette score*	*Calinski Harabasz score*	*Davies Bouldin Score*
40	Master's Program in Regional Language and Literature Education	0.478600304	74.0290282	0.7254216191
41	Master's Program in Indonesian Language and Literature Education	0.5704257078	357.7135453	0.4520157486
42	Master's Program in Physics Education	0.5232653522	855.1331556	0.5248570726
43	Physics Study Program	0.5523061148	1565.782463	0.5020236427
44	Information Systems Program	0.5849994556	126.7535908	0.4350038006
45	Biology Study Program	0.4582402589	4449.41187	0.6466400134
46	Computer Science Program	0.4460617374	1234.057519	0.6570492754
47	Chemistry Study Program	0.5402759459	3901.174183	0.552069023
48	Master's Program in BiologicalSciences	0.7173196489	684.3202966	0.3836404387
49	Master's Program in PhysicalScience	0.6814721065	337.0048772	0.3096593157
50	Master's Program in ChemicalScience	0.7477607198	363.1038902	0.3995511204
51	Master's Program in MathematicalSciences	0.8640376694	382.5221819	0.2228325984
52	Mathematics Study Program	0.5642975938	3243.65023	0.6303855749
53	Agribusiness Study Program	0.5565772378	7450.062786	0.5477881318
54	Master of Forestry ScienceProgram	0.4856209095	235.6277704	0.5907029633
55	Master of Environmental ScienceProgram	0.5318841607	160.8093078	0.5091538986
56	Master's Program in Development Counseling / Community Empowerment	0.8379093817	168.038297	0.2467429631
57	Master's Program in AgriculturalExtension and Communication	0.5355833326	57.33446831	0.380906379
58	Master's Program in AgriculturalIndustrial Technology	0.7088144844	259.8514142	0.3333115891
59	Animal Husbandry Study Program	0.4931679828	515.4441551	0.7053134603
60	Agricultural Engineering Program	0.5249132742	1683.430794	0.5559149788
61	Agricultural Product TechnologyProgram	0.5397172434	4294.245412	0.5226350002
62	Agrotechnology Study Program	0.5088764145	4340.843745	0.6650451961
63	Aquatic Resources Program	0.4743389275	229.4782722	0.5870737257
64	Marine Science Program	0.4465041272	153.5480053	0.7162108935
65	Agricultural IndustrialTechnology Program	0.8567250436	676.9069297	0.2142528295
66	Soil Science Program	0.5467371165	1033.897392	0.5947156204
67	Plant Protection Study Program	0.528859842	663.2001914	0.6909396104
68	Agronomy and Horticulture StudyProgram	0.5051865457	1086.985709	0.5767930031
69	Agricultural Extension Program	0.4592834436	492.188934	0.7261135323
70	Animal Nutrition and FeedTechnology Program	0.5034602914	184.9903997	0.646122724
71	Aquaculture Study Program	0.475145446	404.6190054	0.6790236179
72	Doctoral Program in AgriculturalSciences	0.9449627517	573.7857003	0.05849474941
73	Forestry Study Program	0.4739062901	4254.633823	0.5891520577
74	Master of Agribusiness Program	0.9132297154	315.0793425	0.06293418965
75	Master Program in Agroecotechnology	0.5921873364	280.9192014	0.3916220903
76	Master Program in Agronomy	0.521430067	118.2019792	0.5802355652
77	Geophysical Engineering Program	0.4905781189	5105.186886	0.5259008141
78	Geodetic Engineering StudyProgram	0.4623979648	412.1179729	0.6099492043
79	D3 Civil Engineering StudyProgram	0.6506877125	111.747278	0.3114347
80	D3 Survey and Mapping StudyProgram	0.7639043985	817.3145449	0.3094277958
81	D3 Mechanical Engineering StudyProgram	0.7024992952	122.6001396	0.3374450702
82	Master Program in CivilEngineering	0.6325944622	254.5081069	0.4579902817
83	Master's Program in MechanicalEngineering	0.6359497177	186.8342982	0.333275908
84	Civil Engineering Program	0.5629114663	1678.186894	0.5664636958
85	Electrical Engineering Program	0.4790047524	773.3184727	0.6724525976
86	Chemical Engineering Program	0.6170604178	737.7720447	0.613540253
87	Mechanical Engineering Program	0.4853931879	1343.989841	0.5498384654
88	Informatics Engineering Program	0.6397052667	1076.106831	0.3900090882
89	Architecture Program	0.560260335	339.4545567	0.4816008777

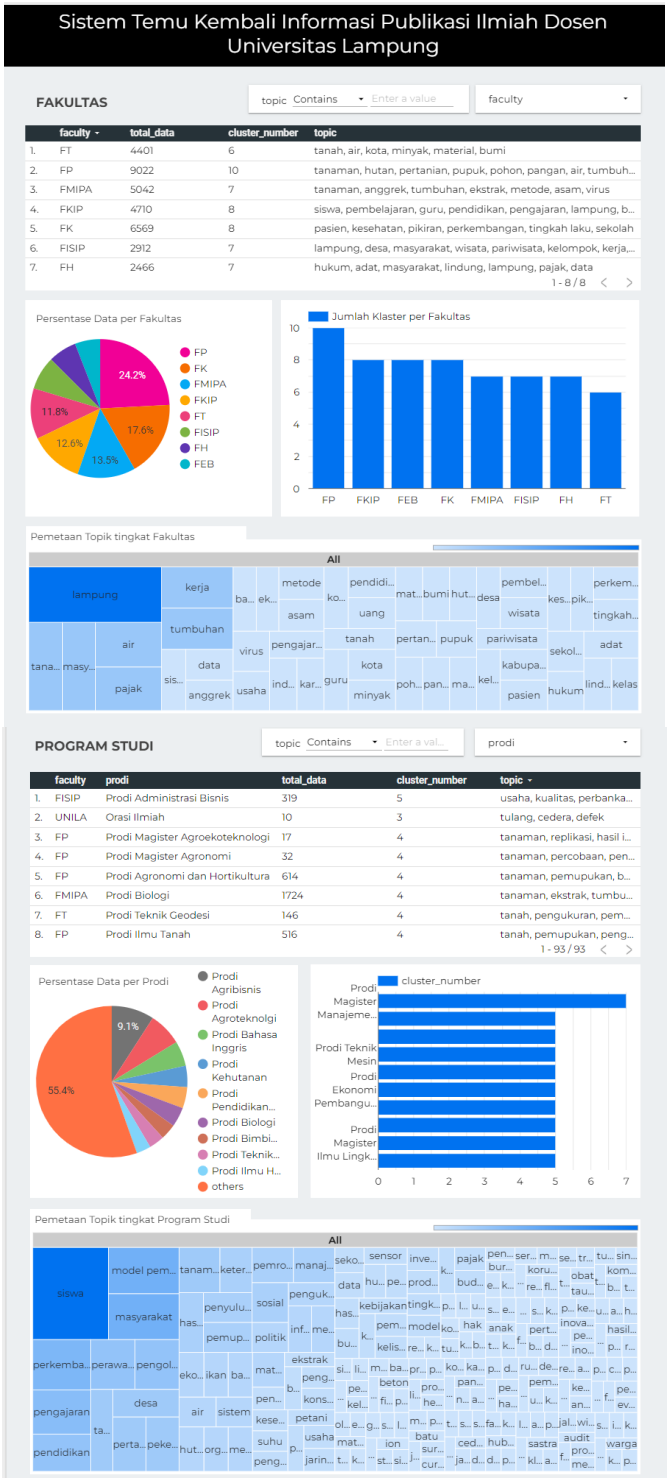

Fig. 6. Dashboard of information retrieval development.

E. Data Distribution

The last stage in the development of this information retrieval system is system distribution. The purpose of this stage is to make it easier for users to access the information generated by the developed information retrieval system.

This system distribution is in the form of a dashboard developed using Google Looker. The information contained in this dashboard consists of general information about this information retrieval system, clustering and classification of topics in the form of detailed data and visualizations. In addition, users can control the menu available on this dashboard and select specific information based on the faculty and study program level. The appearance of this system can be seen in Fig. 6.

F. Analysis Results

From the results of the development of this information retrieval, clustering, topic identification and model evaluation values for each clustering at the University, Faculty and Study Program levels were obtained. For the university level, 7 clusters were obtained with topic classifications including agriculture, soil, education, plants, learning, society, and Lampung. For the faculty level, 6 clusters were obtained for FT, 7 clusters for FMIPA, FISIP, FH, 8 clusters for FKIP, FEB, FK, and 10 clusters for FP.

For the study program level, the most clusters were obtained, namely 5 clusters, this cluster was owned by Master of Government Science study program, Bachelor of Education in Indonesian and Regional Languages and Literature, Bachelor of Education in Drama, Dance and Music, Bachelor of State Administration, Bachelor of Agricultural Engineering, Bachelor of Communication, Bachelor of Agribusiness, Bachelor of Mechanical Engineering, Master of Agricultural Extension and Communication, Diploma in Civil Engineering, Master of Environmental Science, Bachelor of Physical Education, Health and Recreation, Bachelor of Computer Science, Bachelor of Physics. Most of the clusters in this study program belong to study programs with faculties FISIP, FKIP, FT, FMIPA and FP. The amount of data is not always directly proportional to the number of clusters, for example at the faculty level it can be seen that the smallest number of clusters is FT, which is 6, but the amount of data is greater than FEB, FISIP and FT, as well as at the study program level.

Furthermore, the evaluation results of the clustering model obtained for the university level are 0.53 for the Silhouette score, 107298.36 for the Calinski Harabasz score, and 0.60 for the Davies Boulding score. As for the faculty level, the optimal value for Silhouette score is 0.514 which is obtained by FISIP, the optimal value for Calinski Harabasz score is 19386.733 which is obtained by FMIPA, the optimal value of Davies Bouldin score is 0.518 which is obtained by FT. As for the study program level, the optimal value for the Silhouette score is 0.945 which is obtained by the Doctor of Agricultural Science Study Program, the optimal value for the Calinski Harabasz score is 7450.063 which is obtained by the Agribusiness Study Program, the optimal value of the Davies Bouldin score is 0.058 which is obtained by the Doctor of Agricultural Science study program. All study programs obtain the optimal value are come from FP.

In addition, based on the optimal scores obtained from the three clustering evaluation methods used, there are variations in results between faculties and study programs, which indicate the formation of study relationships. Therefore, this study also examined the study relationships between faculties and/or study programs, the results of which can be seen in Table V. FP is the faculty with the highest number of study

relationships, namely 11 relationships. The relationships owned by the Faculty of Agriculture include relationships between study programs in FP, FH, FT, FISIP, and FK.

TABLE V. STUDY RELATIONSHIPS BETWEEN FACULTIES AND STUDY PROGRAMS

Faculty	Faculty relationship	Number of Relationships
FKIP (Faculty of Teacher Training and Education)	[['FKIP1', 'FKIP14'], ['FKIP19', 'FKIP20', 'FKIP21'], ['FKIP23', 'FMIPA3'], ['FK1','FKIP1']]	4
FMIPA (Faculty of Mathematics and Natural Sciences)	[['FMIPA9', 'FT5', 'FT8'], ['FMIPA1', 'FMIPA4'], ['FMIPA6', 'FT6'], ['FISIP1', 'FMIPA5', 'FT5', 'FT8'], ['FKIP23', 'FMIPA3']]]	5
FEB (Faculty of Economics and Business)	[['FEB5', 'FEB6'], ['FEB4', 'FEB6'], ['FEB4', 'FEB5'], ['FEB5', 'FEB2'], ['FEB3', 'FH1','FH3'], ['FEB3', 'FEB6']]]	6
FT (Faculty of Engineering)	[['FMIPA9', 'FT5', 'FT8'], ['FMIPA6', 'FT6'], ['FISIP1', 'FP5', 'FT5', 'FT8', 'FT7'], ['FT9', 'FT12'], ['FISIP1', 'FMIPA5', 'FT5', 'FT8'], ['FP19', 'FP22', 'FT9', 'FT13', 'FT14', 'FT12', 'FT3', 'FT2'], ['FT11', 'FT1'], ['FT11', 'FT4'], ['FT14', 'FP19']]	9
FP (Faculty of Agriculture)	[['FP5', 'FP6', 'FP7', 'FP8', 'FP10', 'FP11', 'FP12', 'FP13', 'FP14', 'FP15', 'FP16', 'FP27', 'FP26', 'FP17', 'FP24', 'FP20', 'FP18'], ['FH1', 'FP1'], ['FP15', 'FP19'], ['FP19', 'FP22'], ['FISIP1', 'FP5', 'FT5', 'FT8', 'FT7'], ['FP1', 'FP25'], ['FP19', 'FP22', 'FT9', 'FT13', 'FT14', 'FT12', 'FT3', 'FT2'], ['FK1', 'FP19', 'FP22'], ['FP1', 'FP25', 'FP2', 'FP24', 'FP19', 'FP22'], ['FP2', 'FP3'], ['FT14', 'FP19'])	11
FISIP (faculty of Social Sciences and Political Sciences)	[['FISIP3', 'FISIP4'],['FISIP1', 'FP5', 'FT5', 'FT8', 'FT7'], ['FISIP1', 'FISIP2', 'FISIP3'], ['FISIP7', 'FK1'], ['FISIP1', 'FMIPA5', 'FT5', 'FT8']]]	5
FK (Faculty of Medical)	[['FKIP1', 'FKIP14'], ['FKIP19', 'FKIP20', 'FKIP21'], ['FISIP7', 'FK1'], ['FKIP23', 'FMIPA3'], ['FK1', 'FP19', 'FP22'], ['FK1', 'FKIP1']]].	6
FH (Faculty of Law)	[['FH1', 'FP1'], ['FEB3', 'FH1', 'FH3']]	2

IV. CONCLUSION

This research aims to develop an information retrieval system that generates clustering and expertise in study fields at the program, faculty, and university levels of Lampung University.

The development stages of this information retrieval system include data collection, data preprocessing, data modeling using VSM, K-Means, and LSA, data evaluation, and system distribution. To address the first research question, VSM is used to obtain text similarity in vector form, K-Means is used for data clustering, and LSA is used to identify study expertise based on the obtained clustering. The analysis results of this information retrieval system development include the number of clusters, the scores from the model evaluation, and the topics identified according to the number of clusters. The obtained cluster numbers for the university level are 7 clusters, while for the faculty level there are 6, 7, 8, and 10 clusters, and for the program level there are 3 to 5 clusters. These cluster numbers answer the second research question. Furthermore, based on the number of clusters, study relationships, and cluster model evaluation scores, study clustering occurs predominantly in programs and faculties of FP, FISIP, FMIPA, FT, and FKIP. Research question number three is addressed by identifying topics based on the number of clusters identified as the most dominant study expertise at Lampung University, including agriculture, soil, education, plants, learning, society, and region Lampung.

Although the topic identification in this research uses Latent Semantic Analysis, the number of identified topics with the number of clusters is still manually selected. Therefore, further research is needed to develop information retrieval systems that can automatically expertise topics based on a set of identification topics. Nevertheless, the development of information retrieval system in this research addresses the needs related to the excellence of study fields at Lampung University.

REFERENCES

[1] Hienert, D., Sawitzki, F., & Mayr, P. (2015). Digital library study in action-supporting information retrieval in sowiport. D-Lib Magazine, 21(3/4), 2015.

[2] Zibani, P., Rajkoomar, M., & Naicker, N. (2022). A systematic review of faculty study repositories at higher education institutions. Digital Library Perspectives, 38(2), 237-248.

[3] Piazzini, T. (2022). Bibliographic control and institutional repositories: welcome to the jungle. Bibliographic control and institutional repositories: welcome to the jungle, 132-142.

[4] Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., & Lin, J. (2022). Making a MIRACL: Multilingual information retrieval across a continuum of languages. arXiv preprint arXiv:2210.09984.

[5] Leticia, T., & Elvis, F. (2014, June). Data mining as a tool for information retrieval in digital institutional repositories. In 3rd International Conference on Computer Science and Service System (pp. 180-183). Atlantis Press.

[6] Angelova, M., Vishnu Manasa, D., Boeva, V., Linde, P., & Lavesson, N. (2018). An Expertise Recommender SystemBased on Data from an Institutional Repository (DiVA). In 22nd edition of the International Conference on ELectronic PUBlishing-Connecting the Knowledge Commons: From Projects to Sustainable Infrastructure, Toronto.

[7] Suma, D. V. (2020). A novel information retrieval system for distributed cloud using hybrid deep fuzzy hashing algorithm. Journal of Information Technology and Digital World, 2(3), 151-160.

[8] Saha, K. K., Ray, S., & Sadhukhan, D. (2022, June). A Lightweight and Precise Information Retrieval System for Organisational Wiki. In International Conference on Frontiers of Intelligent Computing: Theory and Applications (pp. 495-507). Singapore: Springer Nature Singapore.

[9] Vallejo-Huanga, D., Jaime, J., & Andrade, C. (2023, March). Similarity Visualizer Using Natural Language Processing in Academic Documents of the DSpace in Ecuador. In International Conference on Information (pp. 343-359). Cham: Springer Nature Switzerland.

[10] Boukhari, K., & Omri, M. N. (2023). DL-VSM based document indexing approach for information retrieval. Journal of Ambient Intelligence and Humanized Computing, 14(5), 5383-5394.

[11] Malkawi, R., Daradkeh, M., El-Hassan, A., & Petrov, P. (2022). A Semantic Similarity-Based Identification Method for Implicit Citation Functions and Sentiments Information. Information, 13(11), 546.

[12] Abasi, A. K., Khader, A. T., Al-Betar, M. A., Naim, S., Alyasseri, Z. A. A., & Makhadmeh, S. N. (2021). An ensemble topic extraction approach based on optimization clusters using hybrid multi- verse optimizer for scientific publications. Journal of Ambient Intelligence and Humanized Computing, 12, 2765-2801.

[13] Lund, B., & Ma, J. (2021). A review of cluster analysis techniques and their uses in library and information science study: k-means and k-medoids clustering. Performance Measurement and Metrics, 22(3), 161-173.

[14] Joby, D. P. (2020). Expedient information retrieval system for web pages using the natural language modeling. Journal of Artificial Intelligence and Capsule Networks, 2(2), 100-110.

[15] Shahapure, K. R., & Nicholas, C. (2020, October). Cluster quality analysis using silhouette score. In 2020 IEEE 7th international conference on data science and advanced analytics (DSAA) (pp. 747-748). IEEE.

[16] Ashari, I. F., Nugroho, E. D., Baraku, R., Yanda, I. N., & Liwardana, R. (2023). Analysis of Elbow, Silhouette, Davies-Bouldin, Calinski-Harabasz, and Rand-Index Evaluation on K-Means Algorithm for Classifying Flood-Affected Areas in Jakarta. Journal of Applied Informatics and Computing, 7(1), 95-103.

[17] Ashari, I. F., Banjarnahor, R., Farida, D. R., Aisyah, S. P., Dewi, A. P., & Humaya, N. (2022). Application of data mining with the K-means clustering method and Davies Bouldin index for grouping IMDB movies. Journal of Applied Informatics and Computing, 6(1), 07-15.

[18] Pradana, A. W., & Hayaty, M. (2019). The effect of stemming and removal of stopwords on the accuracy of sentiment analysis on indonesian-language texts. Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, 375-380.