5thICITB

Undergraduate Thesis Supervisor Recommendation Based On Text Similarity

Fetty Tri Anggraeny[1], Intan Yuniar Purbasari[2], Eka Fitria Wulandari[3]
[1,2,3]Informatics, Universitas Pembangunan Nasional "Veteran" Jawa Timur
Jl. Raya Rungkut Madya, Surabaya, Indonesia
[1]fettyanggraeny.if@upnjatim.ac.id

ABSTRACT:
The measurement of the similarity of texts is very extensive research. The most common implementation is finding documents that match the search keywords. And the current implementation is to measure the plagiarism of scientific documents. In the academic field, the topic of undergraduate students needs to be done an examination of similarity, in addition to knowing the title of the previous thesis that is similar, also to provide recommendations for suitable supervisors. In this study, we propose a system that can help undergraduate students determine supervisors based on the undergraduate thesis title to be proposed. The experimental results show that the proposed method is quite good as a recommendation system for undergraduate thesis supervisors with an accuracy of 80% in the field of research and 87.5% in the field of lecturer research.

Keywords: Text similarity, Dice Coefficient, Supervisor Recommendation.

1. INTRODUCTION

The undergraduate thesis is the final product of undergraduate students to complete their studies. The thesis is a scientific work written by a bachelor program student who discusses a particular topic or field based on the results of a literature study written by experts, the results of field research, or the results of development (experiments)[4]. According to the *Kamus Besar Bahasa Indonesia*, the thesis is a scientific work written by students as part of academic requirements in Higher Education [8]. The thesis must be taken by students to complete their undergraduate studies. The process includes the submission of a thesis proposal, guidance with lecturers, and thesis examination seminar. The suitability of the students' thesis topic with the lecturer research field is very important, this is to facilitate students in completing their thesis because they have the same knowledge. Students before submitting a topic meet with lecturers who have research topics in accordance with the topic to be submitted. A simple way to do this is to look at the undergraduate thesis record that has been guided by each lecturer or student can asking senior students to ask for advice on which supervisor is suitable with his proposed title. Searching for thesis records manually is not an efficient way, this problem can be solved computerized by measuring the similarity of the proposed thesis text with the approved thesis text record. Based on the results of similarity measurements can be processed as a basis for providing recommendations for the names of lecturers in accordance with the proposed thesis title.

Measurement of text similarity has an important role in document processing, for example searching documents [6]., grouping texts according to the field or type of document [2]. calculating essay values [3]. Plagiarism. [13] text classification (Shereen Albitar, Sebastien Fournier, 2014) and others. Text similarity compares between documents to produce a document similarity value. To measure the similarity of documents, the search for similarity of words is done first. Words can be said to be lexically and semantically similar. Words are lexically similar if they have similar character sequences. Whereas semantically similar if it has the same meaning. For example, the words "tail" and "nail" have a high level of lexical similarity but are not semantic. The words "score" and "value" do not have lexical similarity but semantic. Lexical similarity measurement can be done with string-based, while semantic done with corpus-based and knowledge-based [14].

String similarity measures operate on string sequences and character composition. A string metric is a metric that measures similarity or dissimilarity (distance) between two text strings for approximate string matching or comparison. String similarity is divided into character-based and term-based. Several algorithms included term-based such as Cosine Similarity [1]. Dice Coefficient [7]. Eucliden Distance [2], Jaccard Distance (Shuai Wang, Haoliang Qi, Leilei Kong, 2013) and so on. Dice coefficient gives very good results in validating conference papers on the Conference Management System (CMS). Dice coefficient is used to find out whether the paper submitted is in accordance with the theme of the conference [7]. In the application of text mining, the dice coefficient shows pretty good results equivalent to jaccard distance [1]. Same with research on grouping of common bean which shows that the dice-coefficient and jaccard give the same and very good results [5].

In this research, we proposed undergraduate thesis titles similarity for supervisor recommendation, the similarity will be measured using dice coefficient. The most similar *n*-title are used as the basis for calculating the recommendation score of the supervisor. So the system can be a kind of decision support system for students in selecting supervisors according to the proposed thesis topic.

2. LITERATURE REVIEW

2.1 Text Similarity

Text similarity begins with the text processing stages which include case folding, tokenizing, filtering and stemming. Then it is processed using the Rabin-Karp method which consists of *k*-gram indexing, hashing and string matching stages. The similarity is calculated using the dice coefficient and produces the similarity value of each document in the database and sorted from the largest to the smallest similarity, see Figure 1.

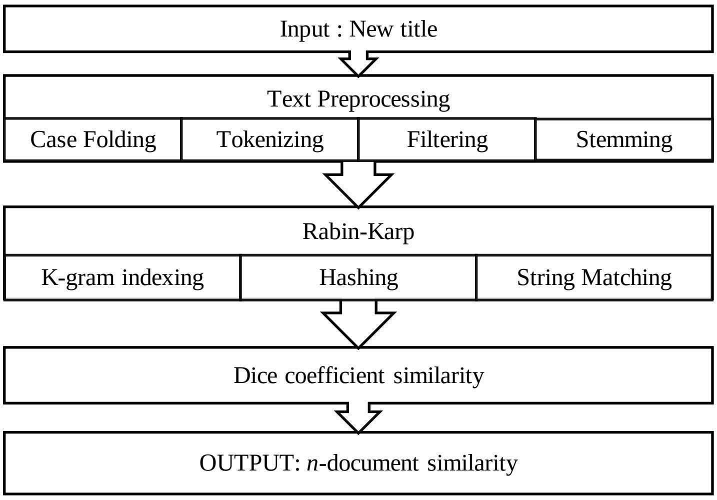

Figure 1: Text similarity.

2.1.1 Text Preprocessing

The text to be performed in the document ranking process generally has several characteristics including having a high dimension, noise in the data, and containing poor text structure. The method used in learning a data text, is by first determining the features that represent each word for each feature in the document. Before determining the features, preprocessing steps are needed, namely case folding, tokenizing, filtering, and stemming.

2.1.1 Rabin-Karp

Rabin-Karp is a search algorithm invented by Michael Rabin and Richard Karp. The Rabin-Karp algorithm can be used to find where a string (in this case called a pattern) is found in another string of larger size, in other words this method uses the hashing function to find the pattern in the text string [9].

2.1.1 Dice Coefficient Similarity

Dice's Similarity Coefficient is used to calculate the similarity value using the k-gram approach (Serhiy Kosinov, 2001). The Similarity value can be calculated using equation 1.

$$S = \frac{2C}{A + B} \qquad (1)$$

Where S is similarity value, A is number of k-gram on text 1, B is number of k-gram on text 2, and C is sum of similar k-gram between text 1 and text 2. To illustrate of dice coefficient, see Table 1 for an example.

Table 1: Example Dice coefficient.

Text 1	Text 2	k	k-gram Text 1	k-gram Text 2	#Similar k-gram	Dice Coefficient
information	informative	2	in-nf-fo-or-rm-ma-at-ti-io-on (10)	in-nf-fo-or-rm-ma-at-ti-iv-ve (10)	8	S = (2*8)/(10+10) = 0.8
information	informative	3	inf-nfo-for-orm-rma-mat-ati-tio-ion (9)	inf-nfo-for-orm-rma-mat-ati-tiv-ive (9)	7	S = (2*7)/(9+9) = 0.78
information	informative	4	info-nfor-form-orma-rmat-mati-atio-tion (8)	info-nfor-form-orma-rmat-mati-ativ-tive (8)	6	S = (2*6)/(8+8) = 0.75

Table 2: Undergraduate thesis data.

#id	Student Name	Title	First Supervisor	Second Supervisor	Research Topic
Doc1	Student 1	Performance Analysis and Simulation of Site to Site IPSec VPN	AJ	CAP	CN
Doc2	Student 2	Identification Of Question And Answer Documents Compatibility Using The Text Mining Method And Vector Space Model	FTA	IYP	CIS
Doc3	Student 3	Making Reseller Recruitment Information System Using The SMART (Simple Multi Attribute Rating Technique) Method (Case Study Of Fresh Water Sepo)	S	RM	SE
Doc4	Student 4	Web-Based Index (Information Security) Application Development	BN	FPA	SE
Doc5	Student 5	Performance Analysis Of Tunneling IP Security (IPSec) And Ethernet Over Internet Protocol (EoIP) On Video Streaming Services	HEW	MI	CN
Doc6	Student 6	Strategic Planning of Information System and Information Technology Using Ward and Peppard	RA	FM	ITSM
Doc7	Student 7	Helthy Food Menu Recommendation System Based On Nutritional Information Using Sugeno Fuzzy Inference System	FTA	YVV	CIS
Doc8	Student 8	Evaluation of Human Resources and Information Technology Using COBIT 4.1	RA	MI	ITSM

2.2 Data

The data used comes from 200 titles the undergraduate thesis data of the Informatics Engineering study program. Each data consist of id, student names, titles, first supervisor, second supervisor, and research topic. Table 2 shows an example of the data and they will used as test data, we used initial for lecturer name. In Informatics UPN "Veteran" Jawa Timur have several research topics are Computer Network (CN), Computing and Intelligent System (CIS), Software Engineering (SE), and Information Technology Service Management (ITSM).

3. METHOD

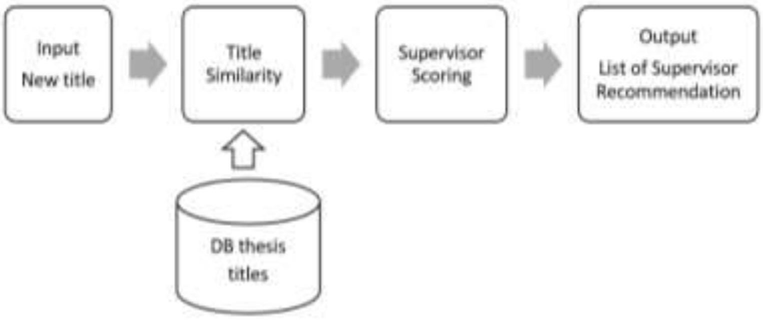

Figure 2: Methodology.

To provide a recommendation for an undergraduate thesis supervisor, the system will perform a title similarity to a new title with the existing thesis titles in the database. After getting n-documents that are similar then the supervisory scoring is carried out to produce a list of supervisor recommendations, see Figure 2). The text similarity method used is the Dice Coefficient with the previous series of information retrieval processes using Rabin-Karp (Figure 1).

The supervisor's scoring is done after the title similarity process is finished and provides n-document similar results. The first supervisor gets more weight scores than the second supervisor, namely a weight ratio of 2:1. Before being used to calculate lecturer scores, similarity scores are normalized first (equation 2). The score of each lecturer is calculated cumulatively based on the supervisor's data from the most similar n-document (equation 3).

$$NS_i = \frac{S_i}{\sum_{i=1}^{n} S_i} \qquad (2)$$

$$LS_j = \sum w * NS_i \qquad (3)$$

For example the previous stage, title similarity, gives the 5-document results most similar to a new title like in Table 3.

Table 3: Example of title similarity results.

New title: *Undergraduate Thesis Supervisor Recommendation Based On Text Similarity (Research Topic = CIS)*					
#document	First supervisor	Second supervisor	Research Topic	Similarity	Normalized Similarity
document-1	A	B	CIS	81.95%	0.29
document -2	C	D	CIS	72.48%	0.26
document -3	E	D	CIS	46.74%	0.16
document -4	F	B	CIS	42.70%	0.15
document -5	B	C	CN	39.80%	0.14
Total				283.67%	1

Based on the five highest similar documents in Table 3, supervisor's score will calculated with the detail shown in Table 4. So the final result of the supervisor's recommendation from the highest to the lowest are B, C, A, D, E, and F.

Table 4: Calculation of supervisor scores.

Lecturer	Supervisor's Score
A	(2 x 0.29) = 0.58
B	(1 x 0.29) + (1 x 0.15) + (2 x 0.14) = 0.72
C	(2 x 0.26) + (1 x 0.15) = 0.67
D	(1 x 0.26) + (1 x 0.16) = 0.42
E	(2 x 0.16) = 0.32
F	(2 x 0.15) = 0.30

4. RESULTS AND DISCUSSION

We used eight test data on Table 2 and use 5-most similar similar to calculate supervisors's score. Table 5, 5-document most similar for the first data give supervisors recommendations are HEW 1.02 and BN 0.62. For the second data in Table 6, FM 0.6 and WSJS 0.45 as supervisors. The third data in Table 7, FTA 0.58 and MSM 0.42 as supervisor, and so on for the rest data. All experiment using parameters *k*=3 and *base*=5 in Rabin-Karp.

Table 5: 5-document most similar with first data.

New title: Performance Analysis and Simulation of Site to Site IPSec VPN					
#document	First supervisor	Second supervisor	Research Topic	Similarity	Normalized Similarity
document-1	HEW	BN	CN	45.36	0.24
document -2	AJ	MI	CN	40.43	0.21
document -3	BN	CAP	CN	36.78	0.19
document -4	HEW	HW	CN	34.95	0.18
document -5	WSJS	HEW	CN	34.23	0.18
Total				191.75	1

Table 6: 5-document most similar with second data.

New title: Identification Of Question And Answer Documents Compatibility Using The Text Mining Method And Vector Space Model					
#document	First supervisor	Second supervisor	Research Topic	Similarity	Normalized Similarity
document-1	WSJS	FM	CN	35.58	0.23
document -2	MI	HEW	CN	30.97	0.20
document -3	CAP	LPP	CIS	30.588	0.20
document -4	BN	HEW	CN	30	0.19
document -5	FM	MSM	SE	29.46	0.19
Total				156.60	1

Table 7: 5-document most similar with third data.

New title: Making Reseller Recruitment Information System Using The SMART (Simple Multi Attribute Rating Technique) Method (Case Study Of Fresh Water Sepo)					
#document	First supervisor	Second supervisor	Research Topic	Similarity	Normalized Similarity
document-1	FM	MSM	CIS	47.76	0.23
document -2	S	RM	SE	41.86	0.20
document -3	FTA	MSM	SE	40.22	0.19
document -4	FPA	MI	ITSM	39.623	0.19
document -5	FTA	YVV	CIS	39.604	0.19
Total				209.07	1.00

A list of recommended lecturers to guide undergraduate thesis for each title can be seen in Table 8. The number of supervisors varies, depending on the 5 most similar documents produced at the title similarity stage.

Table 8: List supervisor each data.

#document	Supervisor
Doc1	HEW, BN, AJ, MI, CAP, HW, WSJS
Doc2	FM, WSJS, MI, CAP, HEW, BN, LPP, MSM
Doc3	FTA, MSM, FPA, FM, S, RM, MI, YVV
Doc4	MSM, YVV, FPA, BN, S, CAP, HEW, RA
Doc5	HEW, WSJS, KR, BN, CAP
Doc6	RA, FPA, MI, BN, FM
Doc7	EYP, RA, YVV, WSJS, MI, FM, MSM
Doc8	RA, FM, FPA, MI

Further, we compare the research topic between data test and n-most similar document, see Table 9. There are two data, Doc2 and Doc7, have wrong research topic. Based on this experiment, the system is quiet good, because has 80% accurate in research topic decision.

Table 9: Supervisor's Score of data.

#document	Target Research Topic	Output Research Topic	Match
Doc1	CN	CN	Yes
Doc2	CIS	CN	No

Doc3	SE	CIS/SE	Yes
Doc4	SE	SE	Yes
Doc5	CN	CN	Yes
Doc6	ITSM	ITSM	Yes
Doc7	CIS	SE	No
Doc8	ITSM	ITSM	Yes

Now, we see the supervisor research field, each supervisor join on one or more research topic, see Table 10. It shows that at least one record is not match and 7 record are match. The system give 87.5% accuration rate in supervisor research topic.

Table 10: Supervisor’s Score of data.

#document	Target Research Topic	Supervisor	Supervisor and Research Topic	Match
Doc1	CN	HEW and BN	HEW: CN and BN: CN, SE, CIS, ITSM	Yes
Doc2	CIS	FM and WSJS	FM: ITSM and WSJS: ITSM, CIS, CN	No
Doc3	SE	FTA and MSM	FTA: SE, CIS and MSM: SE	Yes
Doc4	SE	MSM and YVV	MSM: SE and YVV: CIS	Yes
Doc5	CN	HEW and WSJS	HEW: CN and WSJS: ITSM, CIS, CN	Yes
Doc6	ITSM	RA and FPA	RA: ITSM and FPA: ITSM	Yes
Doc7	CIS	EYP and RA	EYP: CIS and RA: ITSM	Yes
Doc8	ITSM	RA and FM	RA: ITSM and FM: ITSM	Yes

5. CONCLUSIONS

Based on the discussion above, it can be concluded that the recommendation system for thesis supervisors is quite good. It has good accuracy both on the research topic and the supervisor's research topic, 80% and 87.5%. The proposed method can be used as a feature in the Final Assignment Information System application, so students do not experience difficulties in choosing undergraduate thesis supervisors. Furthermore, deeper research can be done to determine the best parameter values, for example *k* and *base*. It is also possible to compare the performance between the dice coefficient and other similarity measurement methods

REFERENCES

[1] Baoli Li, L. H. (2013). Distance weighted cosine similarity measure for text classification. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 611–618). Springer Berlin Heidelberg. https://doi.org/https://doi.org/10.1007/978-3-642-41278-3_74

[2] E. Laxmi Lydia, P. Govindaswamy, SK. Lakshmanaprabu, D. R. (2018). Document Clustering Based on Text Mining K-Means Algorithm Using Euclidean Distance Similarity. *Journal of Advance Research in Dynamical & Control System*, *10*(02),

208–214.

[3] Guntur Budi Herwanto, Yunita Sari, B. N. P., & Mardhani Riasetiawan, Isna Alfi Bustoni, and I. H. (2018). UKARA: A Fast and Simple Automatic Short Answer Scoring System for Bahasa Indonesia. In *International Conference on Educational Assessment and Policy (ICEAP)* (pp. 48–53). https://doi.org/https://doi.org/10.26499/iceap.v2i1.95

[4] Huda, M. (2011). Perkembangan Keilmuan di STAIN Ponorogo. *Dialogia*, *9*(2).

[5] Jair Moura Duarte, João Bosco dos Santos, L. C. M. (1999). Comparison Of Similarity Coefficients Based On Rapdmarkers in The Common Bean. *Genetics and Molecular Biology*, *22*(3), 427–432. https://doi.org/http://dx.doi.org/10.1590/S1415-47571999000300024

[6] Nitesh Pradhan, Manasi Gyanchandani, R. W. (2015). A Review on Text Similarity Technique Used in IR and its Application. *Internation Journal of Computer Application*, *120*(9), 28–34.

[7] Nurhilyana Anuar, A. B. M. S. (2010). Validate conference paper using dice coefficient. *Computer and Information Science*, *3*(3), 139–145.

[8] Poerwadarminta, W. J. S. (2002). *Kamus Umum Bahasa Indonesia*.

[9] Richard M. Karp, M. O. R. (1987). Efficient randomized pattern-matching algorithms. *IBM Journal of Research and Development*, *31*(2).

[10] Serhiy Kosinov. (2001). Evaluation of N-grams Conflation Approach in Text-Based Information Retrieval. *Spire*, 136–142.

[11] Shereen Albitar, Sebastien Fournier, B. E. (2014). An Effective TF/IDF-Based Text-to-Text Semantic Similarity Measure for Text Classification. In *Web Information System Engineering* (pp. 105–114).

[12] Shuai Wang, Haoliang Qi, Leilei Kong, C. N. (2013). Combination of VSM and Jaccard coefficient for external plagiarism detection. In *International Conference on Machine Learning and Cybernetics* (pp. 1880–1885). IEEE. https://doi.org/10.1109/ICMLC.2013.6890902

[13] Stamatatos, E. (2011). Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology*, *62*(12), 2512–2527.

[14] Wael H. Gomaa, A. A. F. (2013). A Survey of Text Similarity Approaches. *International Journal of Computer Applications*, *68*(13), 975–8887.