Journal of Telematics and Informatics (JTI)
Vol.6, No.1, March 2018, pp. 15~20
ISSN: 2303-3703

Indonesian Articles Recommender System Using N-Gram and Tanimoto Coefficient

Supriyanto, Imam Much Ibnu Subroto , Muhammad Khosyi'in
Department of Informatics Engineering, Sultan Agung Islamic University
Jl. Raya Kaligawe KM 4 Semarang 50112, Indonesia
e-mail: supriyanto@stekom.ac.id

Abstract

Human needs of technology and the availability of adequate infrastructure are evidences that the technology currently becomes part of the human beings' basic necessities. Growing multitude of journals and scientific papers makes choosing and sorting become more selective though there have been many online journals service providers and portals. Research on search engines, plagiarism and recommendation system has been carried out with various methods to improve the performance of the system itself, this paper aims to calculate similarities between one article with other articles by implementing n-gram and tanimoto cosine. The number of articles tested were forty-three titles and abstracts, tests were carried fifty times with random selected keywords, by separating each sentences of the title and abstract into n characters (n = 2) including spaces and punctuation, then calculating similarity to the query or keywords used to test the system. Testing was done using several variation of the thresholds. After observing the fifty times-testings, the threshold value of 0.30, produced accuracy = 0.86, precision = 0.37 and recall = 0.44.

Keywords*: Recommendation System, Tanimoto Cosine, Similarity, Accuracy, Precision and Recall*

1. Introduction

Human needs of technology really fasten the development of the technology itself. Rapid technological advances supported by adequate infrastructure and facilities become one of the proofs that technology is becoming a basic requirement for human life. One of the dynamics development of science and technology is information technologies in the field of search engine. Several researches in this field have been done by researchers with various methods with different results. Developing a journal is one of the requirements for students to graduate from universities or other higher education institutions, so there is in need of better management[1], even though there have been many online journal portal service providers.

One result of the abundant information available online is the phenomena of *copy – paste* without mentioning the reference which becomes easy to do and makes the papers categorized as a plagiarism from other scientific works[2]. Research on search engines and a recommendation system with different methods and algorithms has been done before, such as in research[3], which is a full text search done by computer by tracing the whole contents of the document. A recommendation system is designed to predict an object of research[4]. Recommendation system is a program that is able to predict an item, such as recommendations for movies, music, books, news and others which are interesting to users. The system runs based the collected data from a user directly or indirectly using specific algorithms that are considered capable of generating accuracy, precision and recall. The selection of the appropriate algorithms is one of the main requirement in order that the research conducted is able to produce excellent accuracy, precision and recall.

The system is built to calculate the value of similarity between one article or journal with randomly chosen keywords by using *n-gram* method and *cosine tanimoto*, applying threshold is conducted for ease in calculation of accuracy, precision and recall in the analysis process of the test results. *N-gram* is one way to divide the sentence or word into n-characters including

punctuation and spaces, then matching algorithm is conducted using *tanimoto cosine*. *Tanimoto cosine* algorithm is a combination of the *tanimoto* similarity with *cosine* similarity, with the hope of being able to produce better calculation of accuracy, precision and recall compared to using a single algorithm.

2. Related Study

The research reference in this study is the analysis of combined algorithm of weighted tree similarity and tanimoto cosine (tc)[1] the study uses a Weighted Tree Similarity algorithm combined with Cosine Tanimoto algorithm to calculate semantic searching. The next research is semantic searches on the journal portal by detecting the presence of the same sentence as plagiarism indication using n-gram based hashing algorithm[2]. The next research is the application of *weighted tree similarity* algorithm for semantic searching[3] which aims to improve the precision and recall on a search engine. Further research is the development of a recommendation system using the decision tree and clustering[4].

3. Research Method

The research was carried out using research and development method. It was carried out in several stages ranging from collecting data, applying n-grams and using algorithms tanimoto cosine. Tests were done fifty times by applying thresholds of 0.15 0.30 and 0.50, then calculating the value of accuracy, precision and recall of the system built. The following is a description of the carried out research:

3.1. Collecting Data

The data collection begins with collecting and grouping article data along with the abstract with total of forty titles selected randomly, the keywords or query selected were random as well.

Table 1. Research data

No	Title	Keyword
1	usaha bisnis desain grafis	usaha bisnis desain grafis
2	peluang bisnis desain grafis	perancangan sistem informasi
3	desain grafis	dasar desain grafis
4	bisnis desain grafis dan pemasaran	analisa keuangan
5	peluang bisnis di bidang desain grafis	aplikasi perkantoran
6	peluang bisnis industri desain grafis	kelebihan penggunaan sistem
7	peluang bisnis desain grafis dan percetakan	anggaran pendapatan daerah
8	peluang bisnis melalui hobi editing foto (desain grafis)	perancangan animasi
….	….	….
43	citarasa jajan pasar yogyakarta dalam desain grafis	faktor faktor pendukung keputusan

Table 1. is the research data chart, the collection and selection of the article titles and abstracts were selected randomly, total data to be used and would be tested were forty-three with a total of keywords used in the tests were fifty which were also selected randomly.

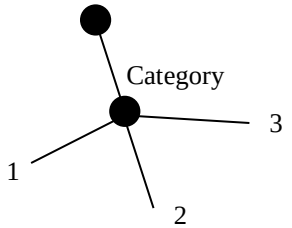

Figure 1. Article grouping

The randomly selected articles then directly grouped into three categories which were graphic design, accounting and computer systems, as in Figure 1 represents a grouping of articles.

3.2. Applying Algorithm

Tanimoto Cosine is a combination of cosine similarity and tanimoto similarity[1], the equation of cosine similarity can be seen in the Equation 1,

$$\sum((S_i) \quad (W_i + W_i)/2) \tag{1}$$

3.3. Designing the System

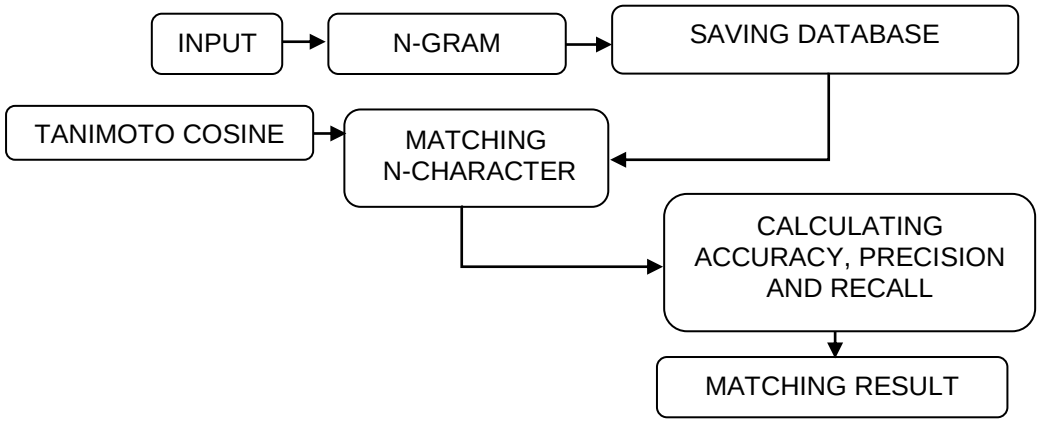

Figure 3. Block diagram of the system design

Designing the system was done in several stages as shown in Figure 3, the collection of randomly chosen articles was grouped into three categories as shown in Figure 1, which were graphic design, accounting and computer system, all the data was divided into n = 2 including spaces and punctuation, then every n characters would be matched by the system with a randomly chosen keyword with the total of fifty. Similarity values would be displayed by the system and sorted from the highest value to the lowest one.

4. Results and Analysis

The tests in the study were conducted fifty times with scambled keywords and resulted in different similarity value.

4.1. Test table

Table 2. Test table

		usaha bisnis desain	perancangan sistem informasi	dasar desain grafis	analisa keuangan	aplikasi perkantoran	kelebihan pengguna an sistem	anggaran pendapatan daerah	perancangan animasi	.....	pengaruh gaji terhadap kinerja karyawan
No	Title	1	2	3	4	5	6	7	8	.....	50
1	usaha bisnis desain grafis	1,00	0,09	0,52	0,09	0,03	0,12	0,15	0,09	.....	0,08
2	peluang bisnis desain grafis	0,70	0,21	0,47	0,21	0,10	0,21	0,25	0,26	.....	0.09

3	desain grafis	0,65	0,12	0,72	0,08	0,03	0,09	0,13	0,08	.....	0,04
4	bisnis desain grafis dan pemasaran	0,65	0,27	0,63	0,14	0,16	0,21	0,34	0,29	.....	0,13
5	peluang bisnis di bidang desain grafis	0,62	0,19	0,45	0,18	0,12	0,19	0,28	0,23	.....	0,12
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....	.....
43	peluang bisnis desain grafis dan percetakan	0,54	0,20	0,39	0,16	0,17	0,20	0,33	0,23	.....	0,30

Table 2 shows the results of similarity value of the test conducted fifty times, the data tested was the title and the abstract with randomly selected keywords. From these results we can see that the up and down of similarity values were influenced by the similarity of the data and keywords.

4.2. Analysis

The tests were carried out using threshold which this study used threshold of 0.50 0.30, and 0.15.

a. Test 1

Table 3. Threshold Test Table

0,50	RELEVANT	NOT RELEVANT
RETRIEVED	7	0
NOT RETRIEVED	3	33

Accuracy = 0,93 Precision = 1 Recall = 0,70

0,30	RELEVANT	NOT RELEVANT
RETRIEVED	10	2
NOT RETRIEVED	0	31

Accuracy = 0,95 Precision = 0,83 Recall = 1

0,15	RELEVANT	NOT RELEVANT
RETRIEVED	10	15
NOT RETRIEVED	0	18

Accuracy = 0,65 Precision = 0,40 Recall = 1

b. Test 2

Table 4. Threshold Test Table

0,50	RELEVANT	NOT RELEVANT
RETRIEVED	0	0
NOT RETRIEVED	0	43

Accuracy = 1 Precision = 0 Recall = 0

0,30	RELEVANT	NOT RELEVANT
RETRIEVED	7	0
NOT RETRIEVED	3	33

Accuracy = 0,93 Precision = 1 Recall = 0,70

0,15	RELEVANT	NOT RELEVANT
RETRIEVED	10	15
NOT RETRIEVED	0	18

Accuracy = 0,65 Precision = 0,40 Recall = 1

c. Test 3

Table 5. Threshold Test Table

0,50	RELEVANT	NOT RELEVANT
RETRIEVED	3	0
NOT RETRIEVED	7	33

Accuracy = 0,84 Precision = 1 Recall = 0,30

0,30	RELEVANT	NOT RELEVANT
RETRIEVED	9	0
NOT RETRIEVED	1	33

Accuracy = 0,98 Precision = 1 Recall = 0,90

0,15	RELEVANT	NOT RELEVANT
RETRIEVED	10	7
NOT RETRIEVED	0	26

Accuracy = 0,84 Precision = 0,59 Recall = 1

Data from the test results can be seen in table 3-5, after observing all the test results average data, it can be seen the comparison of accuracy, precision and recall of the three thresholds used as in Figure 4.

Calculating accuracy, precision and recall refers to the formula of the confusion matrix as in research[5]

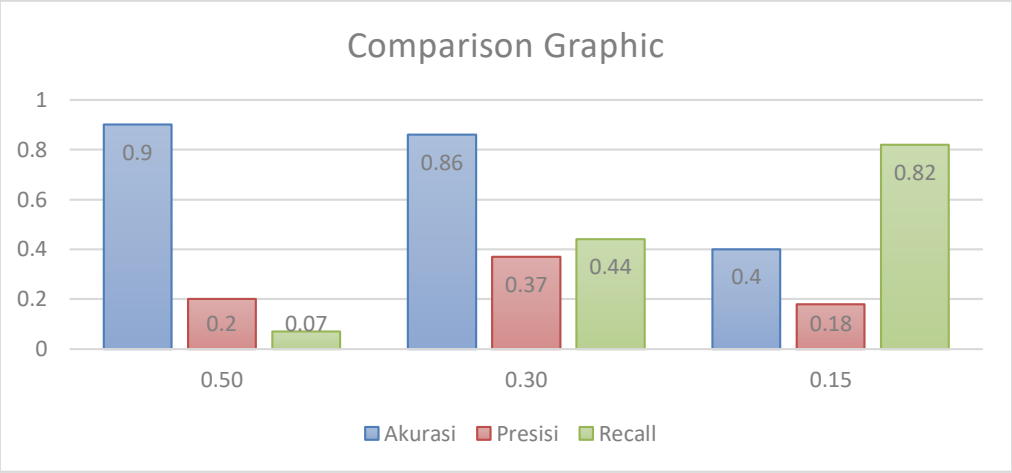

Figure 4. Comparison Graphic

Refer to the threshold comparison chart (Figure 4), it can be seen that the threshold value of 0.50 has the best accuracy with 0.9, precision value of 0.2 and the lowest recall with 0.07, while the threshold of 0.30 is the most excellent in this research refer to the comparison chart in Figure 4. It has 0.37 on the accuracy, 0.86 precision, and 0.44 recall. The threshold 0.15 has accuracy of 0.4, the lowest precision with 0.18 and the highest recall with 0.82.

5. Conclusion

After observing the test results with forty-three article titles and abstracts using fifty random keywords or queries, referring to the threshold table comparison (Figure 4), it can be concluded that the application of n-grams and tanimoto cosine is effective in this study by using a threshold level of 0.30 then it produces the most excellent in terms of accuracy, precision and recall with the value of 0.86 on the accuracy, 0.37 on the precision and 0.44 on the recall.

6. Suggestion

This research immediately divides each sentences into n characters including spaces and punctuation, further research can be developed by clearing the space and punctuation, comparing the results of n = 2, n = 3 and n = 4.

References

[1] Author1 A, Author2 B. Title of Manuscript. *Name of Journal or its Abbreviation*. year; Vol.(Issue): pages.
[2] S. Palgunadi, “Analisis Kombinasi Algoritma Weighted Tree Similarity Dengan Tanimoto Cosine (Tc) Untuk Pencarian Semantik Pada Portal Jurnal,” *Pros. SNST Fak. Tek.*, vol. 1, no. 1, 2014.
[3] D. Purwitasari, P. Y. Kusmawan, and U. L. Yuhana, “Deteksi Keberadaan Kalimat Sama Sebagai Indikasi Penjiplakan Dengan Algoritma Hashing Berbasis N-Gram,” *J. Ilm. KURSOR. Surabaya*, 2011.
[4] R. Sarno and F. Rahutomo, “Penerapan Algoritma Weighted Tree Similarity,” *J. Teknol. Inf.*, vol. 7, no. August, pp. 39–46, 2015.
[5] J. Fadlil and W. F. Mahmudy, “Pembuatan Sistem Rekomendasi menggunakan Decision Tree dan Clustering,” *Age (Omaha).*, vol. 25, no. 26, p. 25, 2007.
[6] C. Matrix, “Confusion Matrix,” pp. 8–10, 2008.