RETRIEVAL INFORMATION IN SEARCHING INTELLIGENTLY ABOUT MUAMALAH

I'lman Teguh Prasetya[1], Galih Hermawan, S.Kom., M.T.[2]

[1,2] Universitas Komputer Indonesia
Jl. Dipatiukur No. 112-116, Bandung 40132
E-mail : ilmanfrasetya@gmail.com [1], galih.hermawan@gmail.com [2]

ABSTRACT

Information Retrieval (IR) is the study of a method for rediscovering relevant information stored in a collection of information sources. In the process of seeking information dibutuhkun a search method. KNN algorithm according to several journals KNN algorithm is considered good in solving case recognition and classification, in this research will be done KNN Algorithm implementation in book search case. In this study to analyze the application of KNN algorithm in book search on the book collection muamalah. The stages are in the analysis phase of preprocessing data. In this study also analyzed at each stage of KNN algorithm that is determination of K, calculation of document testing distance to training document, sort data and determine data testing group. Based on the test results, the KNN algorithm is able to solve the search problem from the information collection. The search process based on word similarity of training data and data testing is very helpful in improving the quality of search results. Search using KNN using 20 training data has 95% accuracy value, search process using KNN algorithm can also give classification information about user problem which make user to be not confusion about material.

***Keywords**: Muamalah, Search using KNN, Information Retrieval System, Classification, Accuracy, Preprocessing Data.*

1. INTRODUCTION

Information Retrieval (IR) is the study of concepts, procedures and methods for rediscovering relevant information stored from various resources or collection of sources of information sought or needed [2]. In the information retrieval process in a source collection, Information Retrieval will require a search method. The K-Nearest Neighbour Algorithm According to Ramdhan [2] has an 80% accuracy in the case of the classification of large task categories [2] and according to Ilyas [17] K-Nearest Neighbors method has 100% accuracy in facial recognition cases. In this research will implement K-Nearest Neighbour Algorithm to solve the search problem.

Muamalah is part of Islamic law that governs the relationship between human beings. In learning activities muamalah need books - books as a reference, At this time a collection of sources of books mumalah already available, but it will be difficult when looking for a book - a book in accordance with the wishes and when doing a search disebuah book collection that has a collection of books that are usually more than One will take a lot of time [3]. This problem raises at least the reference book used or not even found the book sought.

By looking at the above problem, the search for the user needs to go through the preprocessing stage first, then the data that has been processed is searched kemirimpannya value and made a ranking whose purpose shows the book is relevant to the problem entered by the user. After ranking the data is used as a list that will be given to the user. This research uses text minning process to recognize user input and book database, where in the process of searching the basic word (stemming) and using Information Retrieval for withdrawal of information. After the preprocessing phase is completed, the search process of the book and user input data will use K-Nearest Neighbour Algorithm. In this research will be created a system that can help users to find reference books muamalah appropriate user input.

1.1 Problem Identification

Based on the description on the background of the problem, then the formulation of the problem is How to how to find books according to the wishes of the user and provide a list of books required by users in search muamalah book problems.

1.2 Purpose and Objectives

Based on the problems studied, the purpose of this research is to implement K-Nearest Neighbors Algorithm in book search case. While the goal to be achieved from this research know the accuracy of K-Nearest Neighbors Algorithm in the search case.

1.3 Information Retrieval

Information Retrieval handles processes relating to document representation of storage, tracking and discovery of relevant information to a system that can bridge information with users who need information [7]. When performing a search the user inputs the request. This request is similar to the

analysis of the information content of a document, the document is analyzed for the content to know the topic or subject.

1.4 Preprocessing

Preprocessing data is a Phase that encompasses all routines, and processes to prepare data to be used in the operation of knowledge discovery of text mining systems [2]. The purpose of Preprocessing is to simplify the Minning process. In this research Preprocessing stage is done Case Folding, Cleaning Data, Tokenizing, Filtering and Stemming, Stemming using Algorithm Nazief and Andriani, that is algorithm developed for stemming document Indonesian. That is a process of cutting the affixes of a word into its basic word which is an important step in Information Retrieval [1] and Filtering is an important stepping-up phase of Tokenizing [8] .the calculation uses the equation below.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Positives + True\ Positives + False\ Negatives}$$

(1)

1.5 Weighting

Weighting is done to get the value of the word / term successfully extracted. Weight calculation with Term Frequency - Inverse Document Frequency (TF-IDF) uses a combination of two values, ie word frequency and inverse document frequency obtained from dividing the total number of documents as a whole by the number of documents in which the word appears [3].

$$IDF = log(\frac{n}{df}) \quad (2)$$

$$W_{dt} = tf_{df} * IDF_t \quad (3)$$

Description:
n: many collection documents including document testing
d: d document
t: keyword to-t of keyword
W: the weight of the dth document against the t-word
Tf: the number of words searched on a document
IDF: Inversed Document Frequency The IDF value is obtained from IDF: log2 (D / df) where
D: total documents
df: many documents containing the word searched

1.6 K-Nearest Neighbor Algorithm

K-Nearest Neighbor (KNN) is a machine learning method that classifies an object based on learning data closest to the object. The workings of KNN need to involve input in the form of training data, test data and value k. Then sort the train data based on the proximity of the distance based on the count of the data distance tested with the train data. After that it is taken from the top training data k to determine the classification class for the dominant class of the retrieved training data k [10].

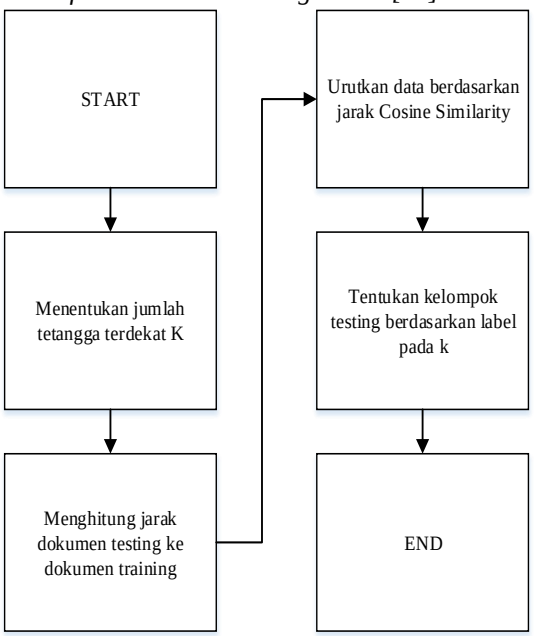


Figure 1. Algorithm flow

1.7 Cosine Similarty

Cosine similarity is a commonality measure commonly used in information retrieval and is the angular measure between the dtest document vectors (dots (i, k)) and dtraining (points (j, k)). Each vector represents every word in every document (text). Cosine similarity formula used is equation 2.3 [10].

$$\cos(\Theta_{ij}) = \frac{\sum_k (d_{ik} d_{jk})}{\sqrt{\sum_k d_{ik}^2} \sqrt{\sum_k d_{jk}^2}} \quad (4)$$

1.8 Confusion Matrix

Confusion matrix is a method that is usually used to perform accuracy calculations on the concept of data mining. Methods using matrix tables [6].

***Table 1.** Table Confusion Matrix*

The correct classification	***Classified as***	
	+	-
+	*True Positives*	*False negatives*
-	*False Positives*	*True Negatives*

True positives is the number of positive records classified as positive, false positives is the

number of negative records classified as positive, false negatives is the number of positive records classified as negative, true negatives is the number of negative records classified as negative, then enter the test data. After the test data is entered into the confusion matrix, compute the values that have been entered to calculate the amount of sensitivity (recall), specificity, precision and accuracy. Sensitivity is used to compare the number of TP to the number of positive records while the specificity is the ratio of the number of TNs to the number of negative records [12]. To calculate used equations

2. CONTENTS OF RESEARCH

At this stage will discuss about the problem to be analyzed. Some of the things contained in this stage is about problem analysis, system analysis to be built and the methods used to solve the problem.

2.1 Problem analysis

Problem Analysis is the stage of identifying problems that will arise upon application of the method. The problem that will be raised in this research is how the application of K-NN method to search the book in accordance with the text in the user input. The search problem is a complex issue because to find the appropriate book requires a suitable approach method to find books that match the user input text.

2.2 System Analysis

System analysis is the stage of explanation and decomposition of the main system into several parts of the system in order to see the needs needed in order to apply the methods used to solve a problem whose purpose is to build an application to perform testing. Applications to be built have several stages. After each stage is done then the output obtained is the search results from user input and tokens that have been given weight. The stages of entering test data and training data books by users. Then the system will process the data. The first step is Case-folding process, Data Cleaning process, Tokenization process, Filtering process, Stemming process and search process using K-NN Algorithm.

2.3 Input Data Analysis

Data input to be analyzed that is training data and test data, data test is user input data in the form of sentence of Indonesian good and true

Table 2. *Table Data Test*

Term
Apakah asuransi dari kantor ABC itu haram ?

Table 3. *Table Data Training*

TITLE	***classification***
Bisnis dan Muamalah Kontemporer	*Kontemporer*

Table 4. *Tabel Data Training*

Number	***TITLE***	***classification***
d1	*Bisnis dan Muamalah Kontemporer*	*kontemporer*
d2	*Bisnis Islami & Kritik atas Praktik Bisnis ala Kapitalis*	*kontemporer*
d3	*Buku Pintar Ekonomi Syariah*	*jual-beli*
d4	*Syariah Marketing (Kertajaya dan Sula)*	*kontemporer*
d5	*Ekonomi Islam Umar bin Khattab*	*jual-beli*

2.4 Preprocessing

Is a process of preparation of test data and training data before the data is processed using KNN Algorithm. Preprocessing itself through several stages include:

Case Folding is the step of changing all data input letters into lower case.

Table 4. *Table implementation Case folding Data Test*

before	***after***
Apakah asuransi dari kantor ABC itu haram ?	*apakah asuransi dari kantor abc itu haram ?*

Tabel 5. *Table implementation case folding Data Test*

before	***after***
Bisnis dan mu'amalah kontemporer tidak bisa dipungkiri telah berkembang, dan menjadi bagian dari kehidupan umat Islam. Namun, tidak semua perkembangan tersebut sesuai dengan syariah. Ketika perkembangan ini menyangkut substansi yang bertentangan dengan syariah. Seperti PT, koperasi, asuransi dan saham. Di sinilah dibutuhkan kecermatan, dan ketepatan pandangan dan sikap. Karena ada juga perkembangan bisnis dan muamalah kontemporer yang dibenarkan oleh syariah. Seperti bisnis online, e-commerce, dan lain-lain. Semuanya bisa ditelaah dalam buku ini. Selamat membaca.	*bisnis dan mu'amalah kontemporer tidak bisa dipungkiri telah berkembang, dan menjadi bagian dari kehidupan umat islam. namun, tidak semua perkembangan tersebut sesuai dengan syariah. ketika perkembangan ini menyangkut substansi yang bertentangan dengan syariah. seperti pt, koperasi, asuransi dan saham. di sinilah dibutuhkan kecermatan, dan ketepatan pandangan dan sikap. karena ada juga perkembangan bisnis dan muamalah kontemporer yang dibenarkan oleh syariah. seperti bisnis online, e-commerce, dan lain-lain. semuanya bisa ditelaah dalam buku ini. selamat membaca.*

Cleaning Data is a step that is used to perform text cleaning of characters other than letters.

Table 6. *Table Data Test*

sebelum	*sesudah*
apakah asuransi dari kantor abc itu haram ?	*apakah asuransi dari kantor abc itu haram*

Tabel 7. *Table Data Test*

before	*after*
bisnis dan mu'amalah kontemporer tidak bisa dipungkiri telah berkembang, dan menjadi bagian dari kehidupan umat islam. namun, tidak semua perkembangan tersebut sesuai dengan syariah. ketika perkembangan ini menyangkut substansi yang bertentangan dengan syariah. seperti pt, koperasi, asuransi dan saham. di sinilah dibutuhkan kecermatan, dan ketepatan pandangan dan sikap. karena ada juga perkembangan bisnis dan muamalah kontemporer yang dibenarkan oleh syariah. seperti bisnis online, e-commerce, dan lain-lain. semuanya bisa ditelaah dalam buku ini. selamat membaca.	*bisnis dan muamalah kontemporer tidak bisa dipungkiri telah berkembang dan menjadi bagian dari kehidupan umat islam namun tidak semua perkembangan tersebut sesuai dengan syariah ketika perkembangan ini menyangkut substansi yang bertentangan dengan syariah seperti pt koperasi asuransi dan saham di sinilah dibutuhkan kecermatan dan ketepatan pandangan dan sikap karena ada juga perkembangan bisnis dan muamalah kontemporer yang dibenarkan oleh syariah seperti bisnis online e-commerce dan lain lain semuanya bisa ditelaah dalam buku ini selamat membaca*

2.4 K-nearest Neighbor Algorithm Analysis

The algorithm analysis is the process of describing or explaining the stages of the K-Nearest Neighbor Algorithm (K-NN or KNN). the stages of Determining the parameters k (the number of nearest neighbors), Calculating the square of the object's euclide distance to the given training data, Sorting the eucliden calculation by ascending, Collecting the Y category (Nearest neighbor classification by k value), Using the nearest most major neighbor category then the object category can be predicted. After all the stages then the results of all the process of classification of the text category into which category.

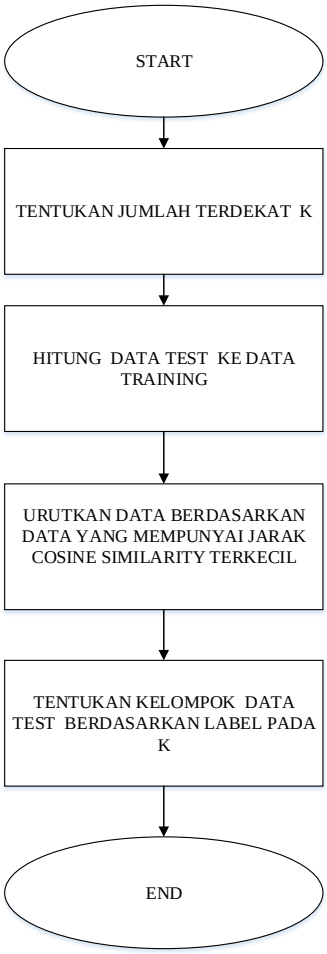


Figure 2. *Stages of K-Nearest Neighbor Algorithm*

2.5 Define Parameter K

Parmeter K is the closest neighboring data parameter. In this process is to enter the number of k in which this number will determine a lot of data that will be selected in the process of grouping and labeling. K disni has been set as many as 3.

2.6 Calculates the Cosine Similarty of test and data training

At this stage perform the Cosine Similiarity calculation process. calculations using data testing and training data that have been weighted using equation 3.

Table 8. *Scalar Multiplication Between Q and other documents*

WDQ.WDI				
D1	***D2***	***D3***	***D4***	***D5***
0	*0*	*0*	*0*	*0*
0	*0*	*0*	*0*	*0*
0	*0*	*0*	*0*	*0*
0	*0*	*0*	*0*	*0*
…	…	…	…	…
WDQ.WDI				
D1	***D2***	***D3***	***D4***	***D5***
0,0062697	*0*	*0,0062697*	*0,0062697*	*0,0062697*
…	…	…	…	…
0	*0,1812381*	*0*	*0*	*0*
0,00627	***0,181238***	***0,00627***	***0,00627***	***0,00627***

Second Calculate the length of each document, including Q. By squaring the weight of each term in each document then add the squared and last value to the sum

Table 9. *Calculate the length of each document Vector*

Length of vector					
Q	***D1***	***D2***	***D3***	***D4***	***D5***
0	*0*	*0*	*0,605194*	*0*	*0*
0	*0*	*0*	*0*	*0,2276447*	*0,2276447*
0	*0*	*0*	*0*	*0,2276447*	*0,2276447*
0	*0*	*0,764874*	*0,015609*	*0*	*0*
0	*0*	*0*	*0*	*0,2276447*	*0,2276447*
0	*0*	*0*	*0,910578*	*0*	*0*
0	*0*	*0,605194*	*0*	*0*	*0*
0	*0*	*0*	*0*	*0,605194*	*0*
0	*0*	*0*	*0*	*0,605194*	*0*
0	*0*	*0,605194*	*0*	*0*	*0*
0	*0*	*0*	*0*	*0,056427*	*0,025078*
0	*0*	*0*	*0*	*0,605194*	*0*
0	*0*	*0*	*0,605194*	*0*	*0*

	Length of vector					
	Q	**D1**	**D2**	**D3**	**D4**	**D5**
	0	0	0	0	0,605519 4	0
	0	0	0	0	0	0
	…	…	…	…	…	…
	0	0	0	0	0,227644 7	0,227644 7
	0	0	0	0,090619 1	0,090619 1	0,090619 1
total	0,702408 1	7,209008	42,74507 7	56,93837 9	115,84414	23,266733
akar	0,838097 9	2,684959 6	6,537971 9	7,545752 4	10,76309 2	4,823560 2

Using equation 3 cosine similarity formula. Calculate the similarity of Q with D1, D2, D3, D4 and D5

Table 10. *Calculation of Cosine Similarty*

Cosine Similarty				
D1	*D2*	*D3*	*D4*	*D5*
0,020086	*0,048909*	*0,0564485*	*0,080517*	*0,036084*

Table 11. *Calculation of Cosine Similarty*

Cosine similarty	**number**	**TITLE**	**classification**
0,020086	*D1*	*Bisnis dan Muamalah Kontemporer*	*kontemporer*
0,036084	*D5*	*Ekonomi Islam Umar bin Khattab*	*jual-beli*
0,048909	*D2*	*Bisnis Islami & Kritik atas Praktik Bisnis ala Kapitalis*	*kontemporer*
0,056449	*D3*	*Buku Pintar Ekonomi Syariah*	*jual-beli*
0,080517	*D4*	*Syariah Marketing (Kertajaya dan Sula)*	*kontemporer*

2.7 Sort data by distance

This process is the process of sorting data based on the value of Cosine Similarty Descending. For the process can be seen in the following table:

Table 12. *Table Implimentation Sort*

Cosine similarty	**NUMBER**	**TITTLE**	**classification**
0,080517	*D4*	*Syariah Marketing (Kertajaya dan Sula)*	*kontemporer*
0,056449	*D3*	*Buku Pintar Ekonomi Syariah*	*jual-beli*
0,048909	*D2*	*Bisnis Islami & Kritik atas Praktik Bisnis ala Kapitalis*	*kontemporer*
0,036084	*D5*	*Ekonomi Islam Umar bin Khattab*	*jual-beli*
0,020086	*D1*	*Bisnis dan Muamalah Kontemporer*	*kontemporer*

2.8 Rank

This process is a ranking process based on the value of Cosine Similarty that has been sorted in descending order.

Table 13. *Implementation Rank*

Ranking	**Cosine similarty**	**number**	**TITTLE**	**classification**
1	*0,080517*	*D4*	*Syariah Marketing (Kertajaya dan Sula)*	*kontemporer*
2	*0,056449*	*D3*	*Buku Pintar Ekonomi Syariah*	*jual-beli*
3	*0,048909*	*D2*	*Bisnis Islami & Kritik atas Praktik Bisnis ala Kapitalis*	*kontemporer*
4	*0,036084*	*D5*	*Ekonomi Islam Umar bin Khattab*	*jual-beli*
5	*0,020086*	*D1*	*Bisnis dan Muamalah Kontemporer*	*kontemporer*

2.9 grouping

With the number k = 3, then the top 3 is taken. Where the top 3 order D4 has a contemporary classification, D3 has a trading classification, contemporary D2 classification.

Table 13. *Implementation grouping*

Ranking	***Cosine similarity***	***number***	***Tittle***	***classification***
1	*0,080517*	*D4*	*Syariah Marketing (Kertajaya dan Sula)*	*kontemporer*
2	*0,056449*	*D3*	*Buku Pintar Ekonomi Syariah*	*jual-beli*
3	*0,048909*	*D2*	*Bisnis Islami & Kritik atas Praktik Bisnis ala Kapitalis*	*kontemporer*

After the data testing (Q) is tested, k the object closest to the data testing is more than the contemporary classification. So Q is classified as a contemporary classification and the recommended book for users is D4 with Shariah Marketing (Kertajaya and Sula), D3 Smart Book of Kertajaya Syariah Economics and Sula and D2 Islamic Business & Criticism on Business Practice ala Capitalist.

Table 14. *Implemenasi detail grouping*

Ranking	***Cosine similarity***	***number***	***Tittle***	***classification***
1	*0,080517*	*D4*	*Syariah Marketing (Kertajaya dan Sula)*	*konteporer*
2	*0,056449*	*D3*	*Buku Pintar Ekonomi Syariah*	*jual-beli*
3	*0,048909*	*D2*	*Bisnis Islami & Kritik atas Praktik Bisnis ala Kapitalis*	*konteporer*

2.9 Testing Measurement Accuracy

Confucion matrix is used for the measurement of classification effectiveness. Using equation 3, an accuracy calculation process was performed for 20 test data.

$$Accuracy = \frac{3+16}{3+1+16+0} \times 100\% = 95\%$$

Based on the test results obtained accuracy to solve the problem of search with data as much as 20 data for 0.95 and can be concluded the accuracy of KNN algorithm will have high value or low accuracy of a training data depending on the quality of training data created.

3. CONCLUSION

3.1 Conclusion

Conclusion Based on the results obtained in this study, it was found that the functionality of the search went well besides KNN accuracy testing algorithm accuracy 95%, But from the search results recommended for the user's book still rank.

3.2 Suggestions

Based on the results of research that has been done, still need to do some further study. As for suggestions for further research are:

1. *Using training data that has categories other than buying and selling and contemporary, so as to achieve a solution that is rich in information and get the optimal solution.*
2. *Optimize the search process to make it faster by using the DBMS system to solve many search problems with data.*
3. *Do the search countermeasures, so that the proposed book is fully optimum.*

BIBLIOGRAPHY

[1] Abdurrasyid, S. A., & Suyanto. (2012). " Implementasi dan Optimasi Algoritma Nazief dan Andriani Untuk Stemming Dokumen Bahasa Indonesia ". Skripsi S1, Fakultas Teknik Informatika, Telkom University.

[2] Dr. Eng R.H. Sianipar, S. M. (2013). " JAVA teori, algoritma, dan aplikasi ". Yogyakarta: ANDI..

[3] Hamilton, K., & Miles, R. (2006). " Learning UML 2.0 ". O'Reilly.

[4] Haryanto, & Priyatmoko, B. (2015). " Sistem Informasi Pengolahan Data Pasien Rawat Inap Berbasis SMS gateway ". ISSN 2337 - 5213 , INFORMATIKA Vol 2. No 1 .

[5] Indriani, A. (2014). " Klasifikasi Data Forum dengan menggunakan ". ISSN: 1907 - 5022, Seminar Nasional Aplikasi Teknologi Informasi (SNATI).

[6] Leidiyana, H. (2013). " Penerapan Algoritma K-Nearst Neighbor untuk penentukan resiko kredit kepemilikan kendaraan bermotor ". Jurnal Penelitian Ilmu Komputer, System Embedded & Logic, 65-76.

[7] Pattah , S. H. (2013). " Peranan Intermediary dalam Sistem Temu Balik Informasi ". ISSN : 2353-9629, Vol. 1 No. 2.

[8] Putri, P. A., Ridok, A., & Indriati. (2016). " Implementasi Metode Improved K-Nearest Neighbor pada Analisis Sentimen Twitter berbahasa Indonesia ". Skripsi S1, Program Teknologi Informasi dan Ilmu Komputer, Universitas Brawijaya.

[9] Rustiana, D., & Rahayu, N. (2017). " Analisis Sentimen Pasar Otomotif Mobil : Tweet Twitter Menggunakan Naive Bayes ". ISSN: 2252-4983, Jurnal SIMETRIS, Vol 8 No 1.

[10] Sani, R. R., Zeniarja, J., & Luthfiarta, A. (2016). " Penerapan Algoritma K-nearest Neighbor pada Information Retrieval dalam penentuan Referensi Tugas Akhir ". Journal of Applied Intelligent System, Vol. 1, No. 2,.

[11] Susandi, D., & Sholahudin, U. (2016). " Pemanfaatan Vector Space Model pada Penerapan ". ISSN: 2406-7741, Jurnal ProTekInfo Vol. 3 No. 1.