TF-IDF ALGORITHM FOR WEIGHTING IN DETERMINING THE SIMILARITY OF TEXT IN DOCUMENTS

Bustami[1]
[1] Doctoral Student of Mathematics and Applied Science, Syiah Kuala University
*Corresponding Author: busabiel@gmail.com

ABSTRACT

The grouping of research documents is needed to facilitate information retrieval. Sometimes we have to read one by one the contents of a document to be able to group it or know the existing information. This research attempts to help in finding information that exists in documents quickly. The information searching in documents by calculating the Term Frequency (TF) and Inverse Document Frequency (IDF) values on each token (word) in each document. The TF-IDF algorithm is an algorithm to calculate the weight of each word that is most commonly used in information retrieval. This algorithm is also known to be efficient, easy and accurate to get results. The accuracy of this algorithm in finding the information in a document reaches above 83,3%.

KEY WORDS: Text mining (Information retrieval), Term Frequency-Inverse Document Frequency (TF-IDF)

INTRODUCTION

The development of Information Technology that is needed by users makes the rapid development of digital documents. These conditions cause some problems to access the desired information accurately and quickly, until the branch of Information Retrieval begins to turn up. Information Retrieval (IR) is the art and science of searching for information in documents, searching for the document itself, searching for metadata that describes the documents, or searching in databases, whether relational databases themselves or hypertext database networks such as the Internet or intranet, for text, voice, image, video or data.

Word detection as a source of information in documents is a way to get a solution in information retrieval (Information Retrieval). From this emerging problem, then a method is developed to deal with the above problems, in this case, the example is the detection of the contents of the research document. The detection process is consists of 2 stages: (1) Document Preparation. At this stage, text manipulation will be carried out on the document which will then be represented in a certain weight in each document (2) Document Classification. For document classification, a threshold value is needed. To get a threshold value (Threshold value) required training data (restros-active document).

In the process of detecting words in documents, a special algorithm is needed to determine the level of similarity between documents based on term composition. In this case, the Term Frequency (TF) and Inverse Document Frequency (IDF) algorithms are needed to calculate the weight for each token (word) contained in the document. After the weight of each document is known, a sorting process is carried out where the greater the value, the greater the level of similarity of the document to the keyword.

LITERATURE REVIEW

Information Retrieval

Information Retrieval (IR) is the science of information retrieval from a number of data that has been lost due to too much data. This knowledge was popularized by Vannevar Bush (1945) and its implementation began to be introduced in the 1950s. In the 1990s, many techniques and methods of retrieval information were developed and used.

The information retrieval system is an activity that aims to provide and supply information to users in response to requests or based on user needs. Basically the information retrieval system is a process to identify, then retrieve a document from a deposit, in response to an information request.

Text Mining

Text mining can be broadly defined as a process of digging up information where a user interacts with a set of documents using analytical tools which are components in data mining, one of which is categorization. Text mining can be considered a relatively new research subject. Text mining can provide solutions to problems such as processing, organizing / grouping and analyzing large amounts of unstructured text.

Tokenisation

Tokenisation is the stage of breaking a set of characters in a text into word units. A set of characters can be whitespace characters, such as enter, tabulation, space. But for single quotes ("), dots (.), Semicolon (;), colon (:) or other, can also have quite a number of roles as word separators. A point (.) Is usually for the end of a sentence, but can also appear in abbreviations, initials, internet addresses, etc. Then a hyphen (-) sign usually appears to combine two different tokens to form a single token. But it can also be found to state the range of values, repetitive words, etc. Or slash character (/) as a file or directory separator or URL or to declare "and or".

Filtering

Filtering is the stopword disposal process which is intended to find out whether a word is entered into stopword or not. Stopword disposal is the process of removing terms that have no meaning or irrelevant. The term obtained from the tokenization stage is checked in a stopword list, if a word is included in the stopword list, the word will enter the next process.

TF-IDF

The TF-IDF method is a method to calculate the weight of each word that is most commonly used in information retrieval. This method is also known to be efficient, easy and has accurate results. This method will calculate the Term Frequency (TF) and Inverse Document Frequency (IDF) values on each token (word) in each document in the corpus. This method will calculate the weight of each token t in document d with the formula:

$$Wdt = tfdt * IDFt$$

Declare:

d = document into d
t = word into t from keyword
W = document weight into d against the word into t
tf = the number of words searched in a document
IDF = *Inversed Document Frequency*

IDF values are obtained from:

IDF = *log2 (D/df)*
D = total document
df = many documents that contain the word you are looking for

After the weight (W) of each document is known, then the sorting process is carried out where the greater the W value, the greater the level of similarity of the document to the keyword.

DISCUSSION

The Information sources that have been tested is the database that has been entered into the database. Then search text data from all existing documents. The author made several experiments for word matching in the articles. After obtaining a matched data on the system with the input text, the author matches the data manually to ensure that data is correct. From the result table, the text mining process carried out by the computer is mostly the same as that done manually. However, there are some articles that do not know the contents of the data because some of the data in the file has the contents of an image so that it cannot be detected by the system. There are the results of the experiments conducted:

No	Kata Kunci Pencarian	Sistem	Manual
1.	PHP umunya digunakan dalam pengembangan web	v	v
2.	PHP merupakana bahasa server-side terpopuler didunia	v	v
3.	PHP mudah di pelajari dan tersedia di semua server	v	v
4.	Hal yang terbaik tentang PHP adalah mudah di gabungkan dengan HTML	v	v
5.	PHP juga merupakan kependekkan dari personal Home Page	v	v
6.	Java adalah sebuah bahasa pemograman yang populer untuk pengembangan perangkat lunak	v	v
7.	Kebanyakan perangkat lunak yang menggunakan java dalah ponsel feature dan smartphone	v	v
8.	Java bahas pemograman multiplatform	v	v
9.	DBMS MySQL	x	v
10.	Java juga termasuk kedalam bahasa pemograman berorientasi objek	v	v
11.	Linux sistem operasi	x	v

	berbasis opensource		
12.	Bahasa pemograman java diciptakan setelah C++	v	v

From articles that have data that matches the keywords entered, a search will be made regarding the amount of data matching so that it can get the highest accuracy and similarity of data that is desired by user.

1. Delete the specific characters

No	Kata Kunci Pencarian
1.	PHP umunya digunakan dalam pengembangan web
2.	PHP merupakana bahasa server-side terpopuler didunia
3.	PHP mudah di pelajari dan tersedia di semua server
4.	Hal yang terbaik tentang PHP adalah mudah di gabungkan dengan HTML
5.	PHP juga merupakan kependekkan dari personal Home Page
6.	Java adalah sebuah bahasa pemograman yang populer untuk pengembangan perangkat lunak
7.	Kebanyakan perangkat lunak yang menggunakan java dalah ponsel feature dan smartphone
8.	Java bahas pemograman multiplatform
9.	DBMS MySQL
10.	Java juga termasuk kedalam bahasa pemograman berorientasi objek
11.	Linux sistem operasi berbasis opensource
12.	Bahasa pemograman java diciptakan setelah C++

2. Set the weight for each term of documents

i = document		c1	c2	c3	c4	c5	m1	m2	m3	m4	m5
all terms		10	7	9	14	9	11	11	4	9	7
1	php	1	1	1	1	1					
2	umumnya	1									
3	gunakan	1			1						
4	dalam	1								1	
5	pengembangan	1					1				
6	web	1									
7	merupakan		1		1	1					
8	bahasa		1		1		1		1	1	1
9	server-side		1								
10	terpopuler	1	1				1				
11	di		1	2	1						1
12	dunia		1								
13	mudah	1		1	1						
14	pelajari			1							
15	dan	1		1				1			
16	tersedia			1							
17	semua	1		1							
18	server			1							
19	hal				1						
20	yang				1		1	1			
21	terbaik				1						
22	tentang				1						
23	adalah				1		1	1			
24	gabungkan				1						
25	dengan				1						
26	html				1						
27	juga					1				1	
28	ke					1				1	
29	pendekkan					1					
30	dari					1					
31	personal					1					
32	home					1					
33	page					1					
34	java						1	1	1	1	1
35	sebuah						1				
36	pemograman						1		1	1	1
37	untuk						1				
38	perangkat						1	1			
39	lunak						1	1			
40	kebanyakan							1			
41	menggunakan							1			
42	ponsel							1			
43	featured							1			
44	smartphone							1			
45	multiplatform								1		
46	termasuk									1	
47	berorientasi									1	
48	objek									1	
49	ciptakan										1
50	setelah										1
51	C++										1

i = document	df(j)	idf(j)
all terms		

1	php	5	0,693
2	umumnya	1	2,303
3	gunakan	2	1,609
4	dalam	2	1,609
5	pengembangan	2	1,609
6	web	1	2,303
7	merupakan	3	1,204
8	bahasa	6	0,511
9	server-side	1	2,303
10	terpopuler	3	1,204
11	di	5	0,693
12	dunia	1	2,303
13	mudah	3	1,204
14	pelajari	1	2,303
15	dan	3	1,204
16	tersedia	1	2,303
17	semua	2	1,609
18	server	1	2,303
19	hal	1	2,303
20	yang	3	1,204
21	terbaik	1	2,303
22	tentang	1	2,303
23	adalah	3	1,204
24	gabungkan	1	2,303
25	dengan	1	2,303
26	html	1	2,303
27	juga	2	1,609
28	ke	2	1,504
29	pendekkan	1	2,303
30	dari	1	2,303
31	personal	1	2,303
32	home	1	2,303
33	page	1	2,303
34	java	5	0,693
35	sebuah	1	2,303
36	pemograman	4	0,916
37	untuk	1	2,303
38	perangkat	2	1,609
39	lunak	2	1,609
40	kebanyakan	1	2,303
41	menggunakan	1	2,303
42	ponsel	1	2,303
43	featured	1	2,303
44	smartphone	1	2,303
45	multiplatform	1	2,303
46	termasuk	1	2,303
47	berorientasi	1	2,303
48	objek	1	2,303
49	ciptakan	1	2,303
50	setelah	1	2,303
51	C++	1	2,303

3. Count Wdt = tf.idf

	c1	**c2**	**c3**	**c4**	**c5**	**m1**	**m2**	**m3**	**m4**	**m5**
1	0,69	0,69	0,69	0,69	0,69	0,00	0,00	0,00	0,00	0,00
2	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
3	1,61	0,00	0,00	1,61	0,00	0,00	0,00	0,00	0,00	0,00
4	1,61	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,61	0,00
5	1,61	0,00	0,00	0,00	0,00	1,61	0,00	0,00	0,00	0,00
6	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
7	0,00	1,20	0,00	1,20	1,20	0,00	0,00	0,00	0,00	0,00
8	0,00	0,51	0,00	0,51	0,00	0,51	0,00	0,51	0,51	0,51
9	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
10	1,20	1,20	0,00	0,00	0,00	1,20	0,00	0,00	0,00	0,00
11	0,00	0,69	1,39	0,69	0,00	0,00	0,00	0,00	0,00	0,69
12	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
13	1,20	0,00	1,20	1,20	0,00	0,00	0,00	0,00	0,00	0,00
14	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00
15	1,20	0,00	1,20	0,00	0,00	0,00	1,20	0,00	0,00	0,00
16	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00
17	1,61	0,00	1,61	0,00	0,00	0,00	0,00	0,00	0,00	0,00
18	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00
19	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00
20	0,00	0,00	0,00	1,20	0,00	1,20	1,20	0,00	0,00	0,00
21	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00
22	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00
23	0,00	0,00	0,00	1,20	0,00	1,20	1,20	0,00	0,00	0,00
24	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00
25	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00
26	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00
27	0,00	0,00	0,00	0,00	1,61	0,00	0,00	0,00	1,61	0,00
28	0,00	0,00	0,00	0,00	1,50	0,00	0,00	0,00	1,50	0,00
29	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00
30	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00
31	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00
32	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00
33	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00
34	0,00	0,00	0,00	0,00	0,00	0,69	0,69	0,69	0,69	0,69
35	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00
36	0,00	0,00	0,00	0,00	0,00	0,92	0,00	0,92	0,92	0,92

37	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00
38	0,00	0,00	0,00	0,00	0,00	1,61	1,61	0,00	0,00	0,00
39	0,00	0,00	0,00	0,00	0,00	1,61	1,61	0,00	0,00	0,00
40	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00
41	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00
42	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00
43	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00
44	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00
45	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00
46	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00
47	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00
48	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00
49	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,30
50	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,30
51	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,30

4. Count the results of the scalar multiplication between Q and other documents

	c1	**c2**	**c3**	**c4**	**c5**	**m1**	**m2**	**m3**	**m4**	**m5**
2	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
6	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
9	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
12	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
14	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00
16	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00
18	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00	0,00
19	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00
21	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00
22	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00
24	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00
25	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00
26	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00	0,00
29	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00
30	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00
31	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00
32	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00
33	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00	0,00
35	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00
37	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00	0,00
38	0,00	0,00	0,00	0,00	0,00	1,61	1,61	0,00	0,00	0,00
40	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00
41	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00
42	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00
43	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00
44	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00	0,00
45	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00	0,00
46	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00
47	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00
48	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,30	0,00
49	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,30
50	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,30
51	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	2,30

	c1	**c2**	**c3**	**c4**	**c5**	**m1**	**m2**	**m3**	**m4**	**m5**
1	0,69	0,69	0,69	0,69	0,69	-	-	-	-	-
2	2,3	-	-	-	-	-	-	-	-	-
3	1,61	-	-	1,61	-	-	-	-	-	-
4	1,61	-	-	-	-	-	-	-	1,61	-
5	1,61	-	-	-	-	1,61	-	-	-	-
6	2,3	-	-	-	-	-	-	-	-	-
7	-	1,2	-	1,2	1,2	-	-	-	-	-
8	-	0,51	-	0,51	-	0,51	-	0,51	0,51	0,51
9	-	2,3	-	-	-	-	-	-	-	-
10	1,2	1,2	-	-	-	1,2	-	-	-	-
11	-	0,69	1,39	0,69	-	-	-	-	-	0,69
12	-	2,3	-	-	-	-	-	-	-	-
13	1,2	-	1,2	1,2	-	-	-	-	-	-
14	-	-	2,3	-	-	-	-	-	-	-
15	1,2	-	1,2	-	-	-	1,2	-	-	-
16	-	-	2,3	-	-	-	-	-	-	-
17	1,61	-	1,61	-	-	-	-	-	-	-
18	-	-	2,3	-	-	-	-	-	-	-
19	-	-	-	2,3	-	-	-	-	-	-
20	-	-	-	1,2	-	1,2	1,2	-	-	-
21	-	-	-	2,3	-	-	-	-	-	-
22	-	-	-	2,3	-	-	-	-	-	-
23	-	-	-	1,2	-	1,2	1,2	-	-	-
24	-	-	-	2,3	-	-	-	-	-	-
25	-	-	-	2,3	-	-	-	-	-	-
26	-	-	-	2,3	-	-	-	-	-	-
27	-	-	-	-	1,61	-	-	-	1,61	-
28	-	-	-	-	1,5	-	-	-	1,5	-
29	-	-	-	-	2,3	-	-	-	-	-
30	-	-	-	-	2,3	-	-	-	-	-
31	-	-	-	-	2,3	-	-	-	-	-
32	-	-	-	-	2,3	-	-	-	-	-
33	-	-	-	-	2,3	-	-	-	-	-
34	-	-	-	-	-	0,69	0,69	0,69	0,69	0,69
35	-	-	-	-	-	2,3	-	-	-	-
36	-	-	-	-	-	0,92	-	0,92	0,92	0,92
37	-	-	-	-	-	2,3	-	-	-	-
38	-	-	-	-	-	1,61	1,61	-	-	-
39	-	-	-	-	-	1,61	1,61	-	-	-
40	-	-	-	-	-	-	2,3	-	-	-
41	-	-	-	-	-	-	2,3	-	-	-
42	-	-	-	-	-	-	2,3	-	-	-
43	-	-	-	-	-	-	2,3	-	-	-
44	-	-	-	-	-	-	2,3	-	-	-
45	-	-	-	-	-	-	-	2,3	-	-
46	-	-	-	-	-	-	-	-	2,3	-
47	-	-	-	-	-	-	-	-	2,3	-
48	-	-	-	-	-	-	-	-	2,3	-
49	-	-	-	-	-	-	-	-	-	2,3
50	-	-	-	-	-	-	-	-	-	2,3
51	-	-	-	-	-	-	-	-	-	2,3

5. Count the length of each document, including Q

6. Apply the cosine similarity formula

$$similarity = \cos(\theta) = \frac{A.B}{||A||\,||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_1^2}}$$

Vector total in A = 92,2170016178487
Vector total in B = 45,8048417819494

$$similarity = \cos(\theta) = \frac{92{,}217.45{,}805}{||92{,}217||\,||B45{,}805||} = \frac{\sum_{i=1}^{n} A92{,}217_1 45{,}805_1}{\sqrt{\sum_{i=1}^{n} A_1^2}\sqrt{\sum_{i=1}^{n} B45{,}805_1^2}}$$

Cos(c, m) = 1

From the text mining search results on the desired data and continued with the calculation of Term Frequency (TF) and Inverse Document Frequency (IDF) values on each token (word), the accuracy of the data found through the TF-IDF algorithm is 83,3%. With the level of similarity of the document to the keyword reaches 1.

CONCLUSIONS

From the results of the research and analysis that has been done, it can be concluded that the use of text mining to get information in the document makes it easy to search. Using the TF-IDF algorithm in weighting search results is very helpful in maximizing accurate information acquisition. From the results of research, the accuracy of the information produced reached 83,3%. From the analysis results also found that the TF-IDF algorithm is very compatible for searching documents with very large numbers and long keywords. But it has a disadvantage if the data in the article is in images, not text.

REFERENCES

Abidin, Taufik Fuadi, Ridha Ferdhiana, and Hajjul Kamil. "Automatic extraction of place entities and sentences containing the date and number of victims of tropical disease incidence from the web." *Journal of Emerging Technologies in Web Intelligence* 5.3 (2013): 302-309.

Aouicha, Mohamed Ben, et al. "Experiments on element and document statistics for xml retrieval." *International Conference on Data, Information and Knowledge Management*. 2008.

Barakbah, Ali Ridho, and K. Arai. "A new algorithm for optimization of K-means clustering with determining maximum distance between centroids." *Proc. Industrial Electronics Seminar (IES)* 2006. 2006.

Fikry, M., Dinata, R. K (2016). Desain Web Dengan HTML dan CSS. *Unimal Press*.

Fikry, Muhammad. "RANCANGAN BASIS DATA KEPENDUDUKAN BERDASARKAN ASPEK-ASPEK KUALITAS SCHEMA DATABASE." *TECHSI-Jurnal Teknik Informatika* 8.2 (2016).

Gil-Leiva, Isidoro. "SISA—Automatic Indexing System for Scientific Articles: Experiments with Location Heuristics Rules Versus TF-IDF Rules." *Knowledge Organization* 44.3 (2017): 139-162.

Hasibuan, Zainal A. "Step-Function Approach for E-Learning Personalization." *Telkomnika* 15.3 (2017).

Ilgisonis, Ekaterina, et al. "Creation of Individual Scientific Concept-Centered Semantic Maps Based on Automated Text-Mining Analysis of PubMed." *Advances in bioinformatics* 2018 (2018).

Maarif, Abdul Azis. "Penerapan Algoritma TF-IDF Untuk Pencarian Karya Ilmiah." *Teknik Informatika Universitas Dian Nuswantoro, Semarang* (2015).

Munadi, Khairul. "INTERACTIVE INTERNET-BASED DISASTER RISK INFORMATION SYSTEM FOR TSUNAMI-HIT ACEH PROVINCE OF INDONESIA." *Journal of Information & Communication Technology* 15.1 (2016).

Savolainen, Reijo. "Pioneering models for information interaction in the context of information seeking and retrieval." *Journal of Documentation* (2018).

Wu, Ho Chung, et al. "Interpreting tf-idf term weights as making relevance decisions." *ACM Transactions on Information Systems (TOIS)* 26.3 (2008): 13.