Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Index group optimization based on automatic clustering using K-Means genetic algorithm

To cite this article: M T Multazam *et al* 2019 *J. Phys.: Conf. Ser.* **1402** 066028

View the article online for updates and enhancements.

Index group optimization based on automatic clustering using K-Means genetic algorithm

M T Multazam[1], R Dijaya[2] and N M S Devi[2]

[1] Law Department, Universitas Muhammadiyah Sidoarjo, Sidoarjo, Indonesia
[2] Computer Science, Universitas Muhammadiyah Sidoarjo, Sidoarjo, Indonesia

*rohman.dijaya@umsida.ac.id

Abstract. E-prints UMSIDA is a repository of student and lecturer publication documents at the Muhammadiyah University of Sidoarjo (UMSIDA). The collection of documents is still random and the search can only detect from the title keywords. The increasing culture of writing and research makes it possible for more and more documents as literature. Documents in E-prints are grouped by subject provided by the repository manager and grouped by the admin who uploaded the document. Automatic document grouping can be done by grouping documents based on the contents of the document using the Information Retrieval (IR) approach. The retrieval process is carried out by document processing with tokenisation to obtain data tokens, the data tokens are processed through a stemming process to obtain the stem value of each word. The stem value is processed using the indexing process and word stem to get sentence indexes through the weighting process. The index results stored in the database become document variables that are the features or characteristics of each document. The index of all documents is grouped through Automatic Clustering technique using K-Means Genetic Algorithm**.**

1. Introduction

Eprints UMSIDA is the storage of student and lecturer publications documents within the Universitas Muhammadiyah Sidoarjo (UMSIDA). The total number of documents in the UMSIDA repository is 2028 documents. Document types from the UMSIDA repository are student publication documents, lecturers, thesis and plagiarism checks that are in the format of *.pdf*, *.jpg,* and *.doc* that have been uploaded by lecturers and students. In searching for documents in UMSIDA the repository only detects the title keywords. Searching for document information in the UMSIDA repository which is limited to the title causes difficulties in finding and grouping documents that are still random. Information Retrieval (IR) is a process that functions to find information that fits the user's needs [1]. The system retrieval information system must be capable of managing imperfect information, and to adapt its behavior to the user context. Information Retrieval aims at defining models and techniques that improves the limitations of current systems for the Information Access [2]. Adopting IR will make it easier for users to search for and obtain the required documents. To achieve this goal, IRs usually use followers process, as in the indexing process, documents are represented in the form of refills, in the screening process, all words stop and general words are filtered and search is the main process of IR [3]. There are several processes in retrieval that will be carried out in this study including tokens, stemming and indexing. Tokenization is the process of separating words from a query based on the sentence that composes them [4]. During tokenization this character stream is separated into tokens, which are the basic units for further processing. Similarly, in applications that receive documents (i.e., character streams representing multiple sentences) as input, the token (or character) stream often needs to be further subdivided to indicate sentence boundaries. The goal of sentence boundary disambiguation is the identification of sentences in such streams. For many applications, tokens correspond roughly to words, but the task of sentence boundary disambiguation can be understood as tokenization, where the desired tokens are sentences [5]. The main purpose of tokenization is to identify words or tokens and the frequency of entering documents [6]. The results of the tokenizing will be stored in the database, then proceed to the next process.

Stemming uses a set of rules has to be applied on a word it will not consider the context of sentence and parts of speech of sentence. In stemming the root is obtaining after applying a set of steps containing set of rules but without taking care off the part of speech (POS) or the occurrence of the context of the word. A variety of stemming algorithms have been developed. Stemming and lemmatizing looks like similar. Both the methods reduce a word variant to its root [7]. The Porter Stemmer: This is one of the most common cutting stemmers are used. This removes the suffix from the word more than a number of iterations to all rules / conditions considered [8]. The Porter algorithm is applied in many fields as a pre-processing step for its indexing tasks. Porter stemmer is actually the most commonly used of all stemmer. The stemmer is based on the idea that suffixes in English are largely constructed from a combination of smaller and simpler suffixes, for example, the suffix "fullness" consisting of two suffixes "full" and "ness". Porter stemmer is a linear stemmer step applied morphological rules in sequence allow gradual removal of affixes. In particular, this algorithm has five steps. Each step defines a set of rules [9]. Such as removing the suffix from words in an automatic way is a special operation useful in the field of information retrieval. In a typical IR environment, someone has collection of documents, each of which is explained by words in the title of the document and possibly words in abstract documents. Ignoring the problem where exactly words come from, we can say that the document is represented by a word vector, or provisions [10]. The stemming process is needed to reduce the number of indexes that are different from a document, besides that the stemming process is also needed to group other words that have basic words and meanings that are similar but have different forms because they get different affixes [11]. Then it will proceed to the indexing process.

The indexing process is a process of getting an index from a collection of documents by using the inverted index technique [12]. An inverted index is a special index which stores the words to bitmap conversion data. In a normal index you would store what words are in a certain document an inverted index is the opposite given a word 'x' what documents have this word is stored. Bitmap index is the index based record numbers to the document [13]. Inverter index using TF-IDF (Term Frequency Inverse Document Frequency) method. TF-IDF is basically a numerical statistic that is meant to show how important a word is to a document in a collection of documents. Information retrieval systems, it is used as a weighting factor. As the number of times a word appears in the document increases, the TF-IDF value increases proportionally. But this TF-IDF value is decreased by the frequency of the word in the collection [14]. The process of IR is tokenization, Stemming, indexing, Automatic Clustering, grouping based on the subject provided. Documents will be processed with tokenization, which is the step of cutting sentences or strings into several words. Then it will be processed using stemming to get the stem value from each word. The stem value is processed using the indexing process of stem words to get the index sentence [15]. The index results of all documents are grouped using Automatic Clustering with K-means Genetic Algorithm by determining document similarity or centroid proximity with index of each data. The cluster data will be in accordance with the groups grouped from the closest distance to the cluster [16].

The process of IR is tokenization, Stemming, indexing, Automatic Clustering, grouping based on the subject provided. Documents will be processed with tokenization, which is the step of cutting sentences or strings into several words. Then it will be processed using stemming to get the stem value from each word. The stem value is processed using the indexing process of stem words to get the index sentence. The index results of all documents are grouped using Automatic Clustering with K-means Genetic

Algorithm by determining document similarity or centroid proximity with index of each data. The cluster data will be in accordance with the groups grouped from the closest distance to the cluster.

2. Methods

2.1. Data description

The data to be processed is a document taken from E-prints UMSIDA at the central of the development of scientific publications (P3I) at the Universitas Muhammadiyah Sidoarjo. The total document in E-prints UMSIDA is 2028 documents. Faculty of Agriculture 40 documents, Faculty of Economics and Business 326 documents, Faculty of Engineering 198 documents, Faculty of Health Sciences 39 documents, Faculty of Islamic Religion 400 documents, Faculty of Law 58 documents, Faculty of Psychology 71 documents, Faculty of Social and Political Sciences 112 documents, Faculty of Science Education and Teacher Training 723 documents, from the point of Muhammadiyah 2 documents, Postgraduate 60 documents. After the data filtering process will be used in the amount of 1000 documents where the type of document that is processed is a document of publication of students, lecturers, thesis in the form of text documents in pdf Indonesian Language. The general process of document content extraction consist of Tokenization, Stemming and Indexing to generate document information as shown in figure 1.

2.2. Tokenisation technique

The tokenisation process is the first stage of the pre-processing process where the sentence will be decapitated into words with spaces as separators. This process will also eliminate punctuation and number characters.

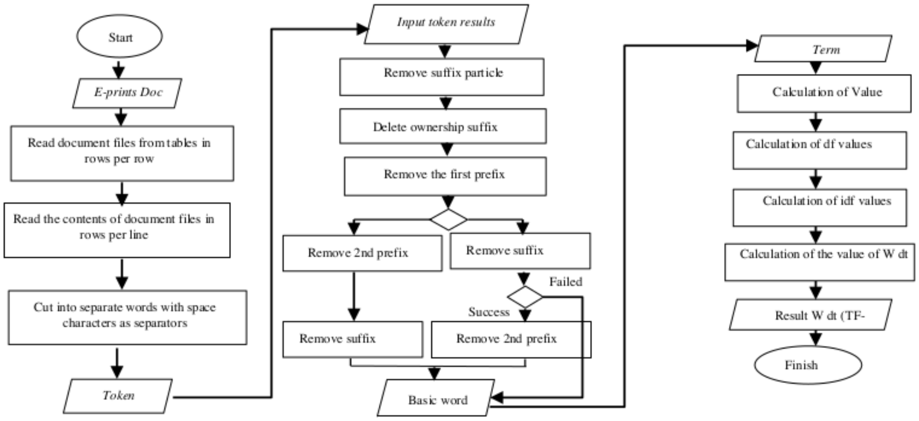


Figure 1. Flowchart document content extraction process.

The steps in the tokenisation process are as follows:

- Upload E-prints documents.
- Then the document files from the table E-prints will be read in rows per line by the system.
- Then the contents of the file from the document will be read in rows per line by the system.
- The system will decapitate into separate words with space characters as separators.
- In the last stage the system of tokenisation will get a token.

2.3. Stemming technique

From the results of the tokenisation process a basic word will be searched by removing the affixes with the pre-processing stemming process. In this study, the stemmer porter algorithm will be used where the basic words that refer to the dictionary will be produced.

The steps in the tokenisation process are as follows:

- The system retrieves the results of tokens that have been previously processed
- Deleting the suffix of a particle
- Removing ownership steps
- Delete the first prefix, if there is no then continue to delete the 2nd prefix, and if it is not found then delete the suffix.
- Erase the 2nd prefix and delete the ending, and the final word is interpreted as a base word.
- Erase the ending if the success process will be continued by deleting the 2nd prefix and the final word is interpreted as the base word, whereas if the process fails it will be immediately interpreted as a base word.
- Deleting a suffix, if it does not exist, the word is interpreted as a base word, but if it is found, it will continue to delete the 2nd prefix and the final word will be interpreted as a base word.

2.4. Indexing technique

From these basic words, it will be processed using an indexing technique to get the index. In this process we will produce word weighting which is calculated using the TFIDF algorithm. Where the index results will be used as search keywords.

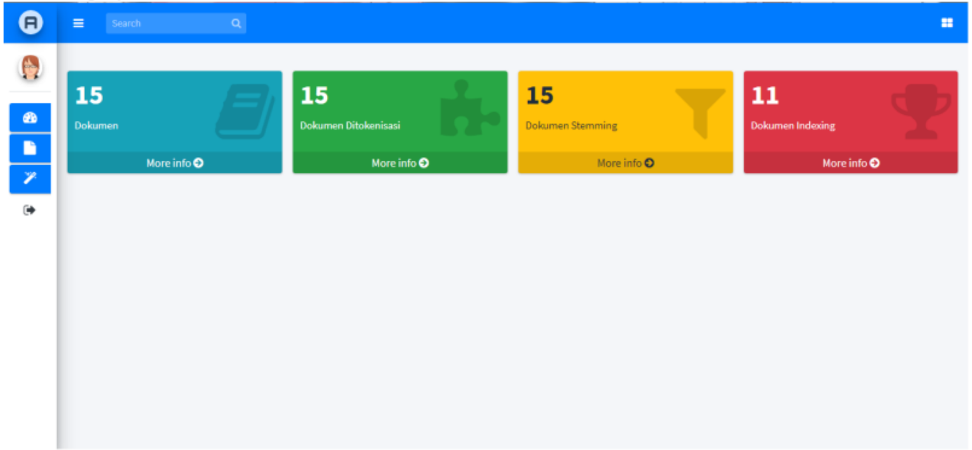
The steps in the indexing process are as follows:

- The system will calculate the term value with the value of DF.
- After getting the DF value, the IDF value will be calculated.
- The system will calculate the value of Weighting.
- Weighting results will be obtained.

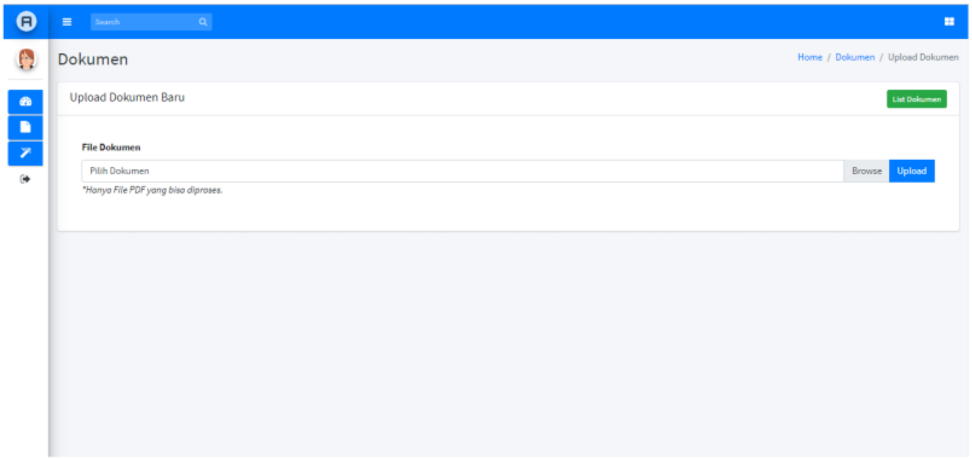
3. Results

In this study a system of index group optimization has been produced using automatic clustering which has been implemented with the PHP and *Laravel* framework programming as shown in figure 2.

3.1. Main page



(a)



Hanya File PDF yang bisa diproses.

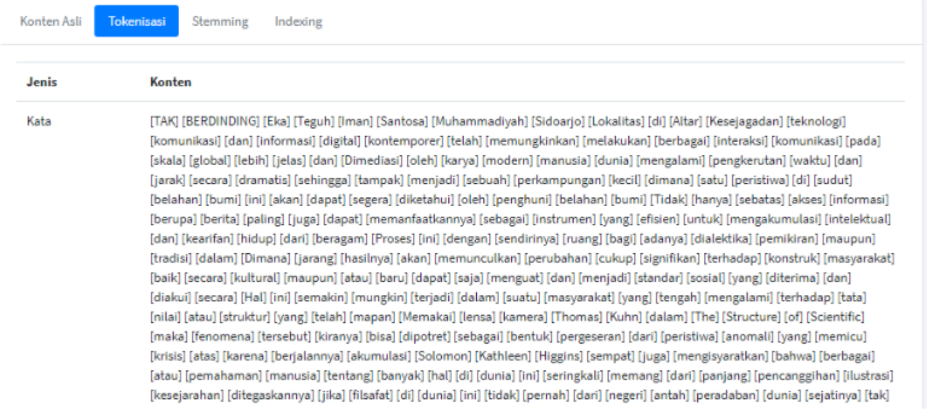
(b)

Figure 2. (a). Document main page and (b). Upload document page.

The main page is the page that will appear first when entering the system. The main page functions to display the homepage of the system. On the homepage it contains all information starting from documents, documents that have been typed, documents that have been installed and documents that have been indexed by clicking the more info feature.

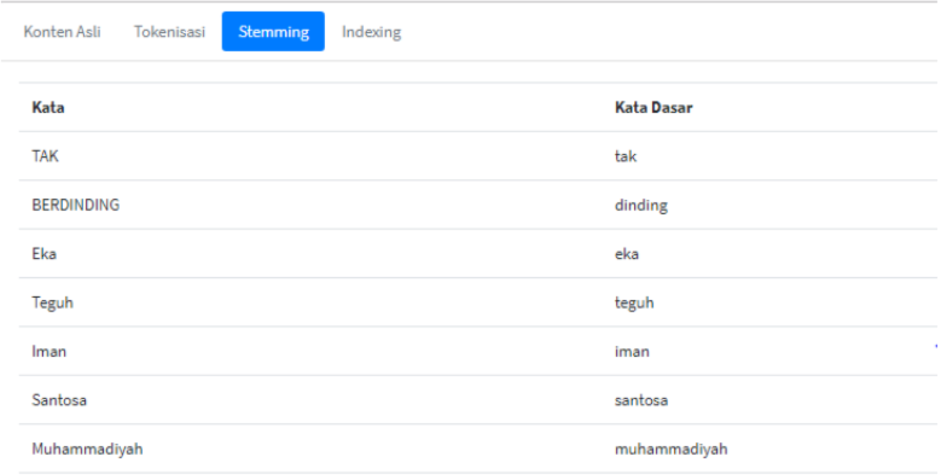
3.2. Upload document page

The document upload page serves as the stage for uploading the E-prints UMSIDA document that has been filtered in the Indonesian .pdf format and will then be saved to the database, after document acquisition on the database the next process is generate token by tokenization process and generate stem by stemming process as shown in figure 3.



Jenis	Konten
Kata	[TAK] [BERDINDING] [Eka] [Teguh] [Iman] [Santosa] [Muhammadiyah] [Sidoarjo] [Lokalitas] [di] [Altar] [Kesejagadan] [teknologi] [komunikasi] [dan] [informasi] [digital] [kontemporer] [telah] [memungkinkan] [melakukan] [berbagai] [interaksi] [komunikasi] [pada] [skala] [global] [lebih] [jelas] [dan] [Dimediasi] [oleh] [karya] [modern] [manusia] [dunia] [mengalami] [pengkerutan] [waktu] [dan] [jarak] [secara] [dramatis] [sehingga] [tampak] [menjadi] [sebuah] [perkampungan] [kecil] [dimana] [satu] [peristiwa] [di] [sudut] [belahan] [bumi] [ini] [akan] [dapat] [segera] [diketahui] [oleh] [penghuni] [belahan] [bumi] [Tidak] [hanya] [sebatas] [akses] [informasi] [berupa] [berita] [paling] [juga] [dapat] [memanfaatkannya] [sebagai] [instrumen] [yang] [efisien] [untuk] [mengakumulasi] [intelektual] [dan] [kearifan] [hidup] [dari] [beragam] [Proses] [ini] [dengan] [sendirinya] [ruang] [bagi] [adanya] [dialektika] [pemikiran] [maupun] [tradisi] [dalam] [Dimana] [jarang] [hasilnya] [akan] [memunculkan] [perubahan] [cukup] [signifikan] [terhadap] [konstruk] [masyarakat] [baik] [secara] [kultural] [maupun] [atau] [baru] [dapat] [saja] [menguat] [dan] [menjadi] [standar] [sosial] [yang] [diterima] [dan] [diakui] [secara] [Hal] [ini] [semakin] [mungkin] [terjadi] [dalam] [suatu] [masyarakat] [yang] [tengah] [mengalami] [terhadap] [tata] [nilai] [atau] [struktur] [yang] [telah] [mapan] [Memakai] [lensa] [kamera] [Thomas] [Kuhn] [dalam] [The] [Structure] [of] [Scientific] [maka] [fenomena] [tersebut] [kiranya] [bisa] [dipotret] [sebagai] [bentuk] [pergeseran] [dari] [peristiwa] [anomali] [yang] [memicu] [krisis] [atas] [karena] [berjalannya] [akumulasi] [Solomon] [Kathleen] [Higgins] [sempat] [juga] [mengisyaratkan] [bahwa] [berbagai] [atau] [pemahaman] [manusia] [tentang] [banyak] [hal] [di] [dunia] [ini] [seringkali] [memang] [dari] [panjang] [pencanggihan] [ilustrasi] [kesejarahan] [ditegaskannya] [jika] [filsafat] [di] [dunia] [ini] [tidak] [pernah] [dari] [negeri] [antah] [peradaban] [dunia] [sejatinya] [tak]

(a)



Kata	Kata Dasar
TAK	tak
BERDINDING	dinding
Eka	eka
Teguh	teguh
Iman	iman
Santosa	santosa
Muhammadiyah	muhammadiyah

(b)

Figure 3. (a) Tokenisation result page and (b) Stemming result page.

3.3. Tokenization pre-processing page

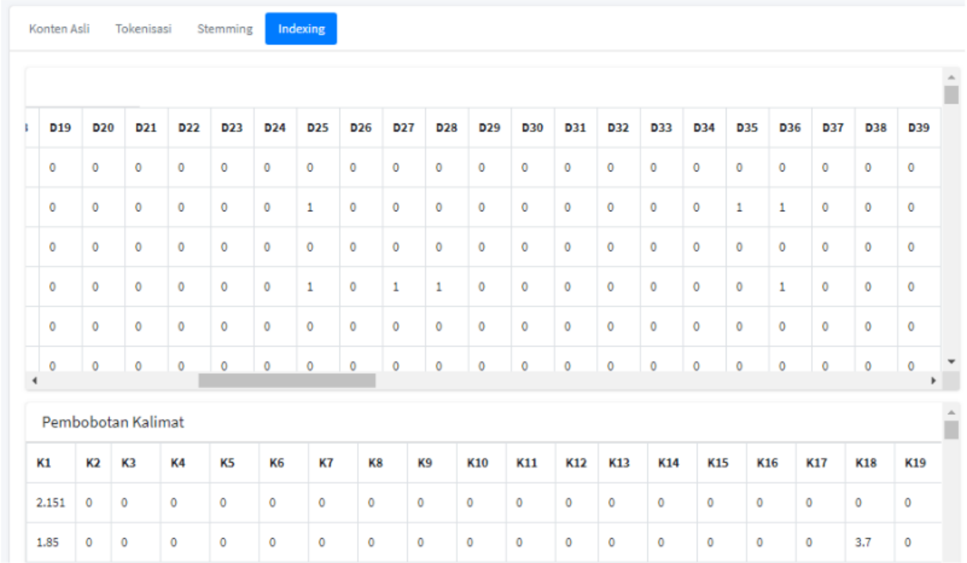
The first preprocessing is the tokenisation process carried out to get a collection of words based on the separator [] to separate words from the sentence. The results of the collection of words will be stored in the token database. Where in this process will be eliminated characters and numbers.

3.4. Pre-processing stemming

The second display of the preprocessing menu is the stemming process. The stemming process functions for the process of changing words into basic words. After getting the basic word from each word, the results will be saved into the database

3.5. Indexing page

The page view of the indexing process is the process of getting an index from a word using the inverted index method. The page will display word weight using TF-IDF, weighting sentences and ranking from weighting results as shown in figure 4.



D19	D20	D21	D22	D23	D24	D25	D26	D27	D28	D29	D30	D31	D32	D33	D34	D35	D36	D37	D38	D39
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Pembobotan Kalimat

K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12	K13	K14	K15	K16	K17	K18	K19
2.151	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1.85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3.7	0

Figure 4. Indexing result page.

4. Conclusion

Document in UMSIDA E-prints Repository are extracted into text, numbers and characters through the tokenisation process. The text from the token process in carried out by the basic search process trough stemming to become a stem. The frequency of each stem in the document is calculated by the weight value using the TF-IDF algorithm. The value of the frequency of the stem in each document will produce on index meta data which can be interpreted as representative of the contents of the document. The index result stored in the database become document variables that are the features or characteristics of each document. The index of all documents is grouped trough Automatic Clustering technique. In the future the results of the cluster can optimize the classification of new documents added to the repository based on the contents of the document.

Acknowledgments

We hereby thank you to Universitas Muhammadiyah Sidoarjo for supporting the publication of this research.

References

[1] Ernawati 2018 Perpustakaan Digital Dalam Temu Kembali Informasi Dengan OPAC *JIPI (Jurnal Ilmu Perpust. dan Informasi)* **3** (1) 103–120
[2] Bansal S and Kumar N 2010 Efficient Information Retrieval Using Document Clustering *_Internasional J. Adv. Res. Comput. Sci.* **1** (3) 22–27
[3] Saini B, Singh V and Kumar S 2014 Information Retrieval Models and Searching Methodologies: Survey *Int. J. Adv. Found. Res. Sci. Eng.* **1** (2) 57–62
[4] Dhony Syafe'i Harjanto N B and Endah S N 2012 Sistem Temu Kembali Informasi pada Dokumen Teks Menggunakan Metode Term Frequency Inverse Document Frequency (TF-IDF) *J. Sains dan Mat.* **20** (3) 64–70
[5] Maletti A 2016 Survey:Finite-state technology in natural language processing *Theor. Comput. Sci.*
[6] Vikram Singh B S 2014 An Effective Pre-Processing Algorithm For Information Retrieval

Systems *Int. J. Database Manag. Syst. (IJDMS)* **6** (6) 13–24

[7] Singh S and Pateriya R K 2015 A Survey on various Stemming Algorithms *Int. J. Comput. Eng. Res. TRENDS* **2** (5) 310–315

[8] Joshi A, Thomas N and Dabhade M 2016 Modified Porter Stemming Algorithm *Int. J. Comput. Sci. Inf. Technol.* **7** (1) 266–269

[9] Karaa W B A 2013 A New Stemmer To Improve Information Retrieval *Int. J. Netw. Secur. Its Appl.* **5** (4) 144–154

[10] Porter M F 2006 An algorithm for suffix stripping *Emerald* **40** (3) 211–217

[11] Wibowo J 2016 Aplikasi Penentuan Kata Dasar dari Kata Berimbuhan pada Kalimat Bahasa Indonesia dengan Algoritma Stemming *J. Ris. Komput.* **3** (5) 346–350

[12] Monica Mayeni A S and Winarno W W 2016 Information Retrieval Dokumen Tesis Untuk Mengetahui Kemiripannya Dengan Penelitian Yang Telah Ada *Transform. J. Inf. Pengemb. Iptek* **12** (2) 105–115

[13] Patil S J and Budhwant D K 2017 Efficient Information Retrieval Using Indexing *Int. J. Comput. Sci. Netw.* **6** (2) 106–109

[14] Savyanavar P A, Mehta B, Marathe V, Padvi P and Shewale M 2016 Multi-Document Summarization Using TF-IDF Algorithm *Int. J. Eng. Comput. Sci.* **5** (4) 16253–16256

[15] Yasid A 2014 Implementasi Automatic Clustering Menggunakan Differential Evolution Dan Cs Measure Untuk Analisis Data Kemahasiswaan *J. Ilm. NERO* **1** (2) 47–52

[16] Afonso A R and Duque C G 2014 Automated Text Clustering Of Newspaper And Scientific Texts In Brazilian Portuguese: Analysis And Comparison Of Methods *JISTEM - J. Inf. Syst. Technol. Manag.* **11** (2) 415–436

Artikel Index Prosiding Scopus.pdf

ORIGINALITY REPORT

16%	11%	14%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Innovations in Information and Communication Technologies (IICT-2020)", Springer Science and Business Media LLC, 2021 Publication	3%
2	ijecs.in Internet Source	3%
3	Submitted to University of Baghdad Student Paper	2%
4	Mahendra K. Ugale, Shweta J. Patil, Vijaya B. Musande. "Document management system: A notion towards paperless office", 2017 1st International Conference on Intelligent Systems and Information Management (ICISIM), 2017 Publication	2%
5	repository.futminna.edu.ng:8080 Internet Source	2%
6	iopscience.iop.org Internet Source	1%

7	www.irjet.net Internet Source	1 %
8	www.emeraldinsight.com Internet Source	1 %
9	www.atlantis-press.com Internet Source	1 %
10	doi.org Internet Source	1 %

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		