Atlantis Highlights in Engineering (AHE), volume 1
International Conference on Science and Technology (ICST 2018)

Document Retrieval System Based on Topic Clustering Method

P.M. Prihatini[1]
Doctoral Engineering Science Department, Udayana University
Electrical Engineering, Politeknik Negeri Bali
Bali, Indonesia
[1]manikprihatini@pnb.ac.id

I.K.G.D. Putra[2]
Information Technology Department
Udayana University
Bali, Indonesia
[2]ikgdarmaputra@unud.ac.id

I.A.D. Giriantari[3], M. Sudarma[4]
Electrical Engineering Department
Udayana University
Bali, Indonesia
[3]dayu.giriantari@unud.ac.id, [4]msudarma@unud.ac.id

***Abstract*—Document retrieval aims to find documents in a collection of unstructured text to meet the needs of user information. The search engine was required in the document retrieval system to perform the entire process automatically, starting from the processing of document text in the collection, feature selection, feature extraction, query text processing and search documents relevant to the query. There were three main factors in improving search engine performance: the feature selection method, the method of weighting features in document collections and the method of searching documents in the collection. In this paper, there were some methods used to improve the performance of search engines. For feature selection, Term Frequency-Invers Document Frequency based on Luhn's Idea was used for document features selection. For weighting features, Fuzzy Gibbs Latent Dirichlet Allocation was used for feature extraction method to weight the document features. To search documents that were relevant to the query, this paper used a Document Retrieval based on Topic Clustering method. Through this method, all documents were clustered based on feature weight obtained through feature extraction methods. Clusters that relevant to the query term combinations were selected and all documents in the cluster were displayed as search results. The result showed this method can retrieve set of documents in the cluster that relevant to the query. Therefore, this method could eliminate the query-document distance calculation function in the retrieval process, so it was hoped that the search process would run faster.**

Keywords— document retrieval; topic model; clustering

I. INTRODUCTION

In completing the work, the user needs information derived from documents in a large collection on the computer through the Information retrieval (IR) process. The classic search model in IR starts from the user's information need which is then changed to a query. The search engine uses these queries to match the relevant documents in the document collection, and then displays the relevant document as a result to the user. If the results do not match the user's needs, the user can repair the query and the process repeated again.

IR requires initial data in the form of unstructured text documents collection, so that it needs to be processed first. Text processing starts from the process of breaking unstructured text into term or tokenization which involves the removal of punctuation or case folding, removal of the most common words or stopwords, and the determination of the root word or stemming. The result of this text processing is the sequence of terms that appear in each document in the collection. The appearance of each term in each document in the collection has different weights, so it requires the next process which is called feature selection and extraction. Furthermore, IR requires user information need of unstructured text, so that it needs to be processed as well so that it has the same format as the document collection. In the classic model, each term in the query is modeled as a boolean, 1 states its existence in the document or 0 for the opposite, then all terms in the query are associated with the logical operator. The boolean value will be used to calculate the similarity distance between queries with each document in the collection using a certain distance function. All documents in the collection will be sorted by the closest distance to the query and displayed as a result of information search. So, there are three main factors in improving search engine performance. First, the feature selection method for determine the important features in the collection. Second, the method of weighting features in document collections and query plays an important role in determining the distance between document-queries. Third, the method of searching documents in collection to speed up the searching process, so it can display documents that are relevant to the query.

Several studies have been conducted with many methods that improve performance of search engines. Hong, Lee, and Han used Term Document Matrix (TDM) as feature selection

method and Genetic Algorithm (GA) as feature extraction method for text clustering and classification [1]. Masoumeh and Seeja used Chi-Square and Information Gain (IG) as feature selection method, Principle Component Analysis (PCA) and Latent Semantic Analysis (LSA) as feature extraction method for text classification [2]. Mahajan and Sharmistha used Wavelet Packet Transform as feature selection method for short text classification [3]. Wang, Zhou, Jin, Liu, and Lu used four methods: One-Hot Encoding, Term Frequency-Invers Document Frequency (TF-IDF) Weighting, word2vec and paragraph2vec which were applied to the short text classification with Naive Bayes (NB), Logistic Regression, Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Decision Tree as classifier [4]. Harish and Revanasiddappa compared the eight methods, TF-IDF, IG, Mutual Information, Chi-Square, Ambiguity Measure, Term Strength, Term Frequency-Relevance Frequency and Symbolic Feature Selection with five different classifiers, NM, KNN, Centroid Based Classifier, SVM and Symbolic Classifier to categorize text documents [5]. Joseph, Mugauri, and Sumathy used four methods as feature, Selection Document Frequency, Standard Deviation Information Gain, Chi Square, and Weighted-Log Likelihood Ratio, for sentiment analysis [6]. Wu, Zhu, Li, Cui, Huang, Li, Chen, and Xu used TF-IDF and Heuristic Selection based on Wikipedia matching approach for text document classification [7]. Prihatini, Putra, Giriantari, and Sudarma used TF-IDF as feature selection method and Fuzzy Gibbs Latent Dirichlet Allocation as feature extraction method for clustering news digital text, and resulted that the topic model gives better results in performing feature extraction than the classical model because the topic model distributes the term into all topics with different probabilities, while the classical model distributes the term only to one topic [8][9].

This paper discusses three main factors in improving search engine performance. TF-IDF based on Luhn's Idea used as a feature selection method; with Fuzzy Gibbs Latent Dirichlet Allocation as a feature extraction method to weight the document features. To search documents that are relevant to the query, this paper uses a retrieval method based on hierarchical clusters. All documents were clustered based on feature weight obtained through feature extraction methods. The retrieval process was done by combining all term with logic function and tracing it to document cluster collection. Clusters that relevant to the query term combinations were selected and all documents in the cluster were displayed as search results. Thus, the query search process eliminated the query-document distance calculation function in the collection, so it is hoped that the search process would run faster. This method becomes the advantages as well as the novelty of this research.

II. Related Research

A. Term Frequency Inverse Document Frequency

The most commonly method for selecting features in document retrieval is Term Frequency Inverse Document Frequency (TF-IDF) [10]. This method selects features based on the number of occurrences of terms in each document as in (1). Variable $tf\text{-}idf_{t,d}$ referred to the TF-IDF value of term t in the document d. Variable $tf_{t,d}$ referred to term frequency value of term t in the document d. Variable N referred to the total number of documents in collection. Variable df_t referred to document frequency value of term t.

$$\text{tf-idf}_{t,d} = \text{tf}_{t,d} \times \log (N/\text{df}_t) \quad (1)$$

B. Luhn's Idea

Luhn's Idea is a concept used to select features based on the TF-IDF value of each term in the document [11]. The number of occurrences of terms in the document is compared with the level of importance of the terms in the document so as to produce a curve. The terms that were above the upper limit value and below the lower limit value were obtained through the curve. Those terms are document features that did not affect the contents of the document, so it was removed from the feature list. The remaining terms between the lower and upper limit values were selected as the selection feature used in the feature extraction stage.

C. Fuzzy Gibbs Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a topic-based feature extraction method at term, document and corpus level [12]. Gibbs Sampling is one of the reasoning methods used for LDA [9]. LDA requires a sampling process that is carried out repeatedly until it reaches convergent conditions.

To improve LDA performance in achieving convergent conditions, the concept of fuzzy logic is added in the LDA sampling process [8] as in (2). Variable $p_{t,k}$ refers to weighted value for term t of topic k. Variable nkw_{-1} refers to the value of the topic-term matrix by ignoring the current term value. Variable V is the unique number of terms in all documents. Variable ndk_{-1} refers to value of the document-topic matrix by ignoring the current term value. Variable β determine the mixing proportion of documents on the topics, while α determines the mixture components of words on the topics [13]. Variable K is the number of topic.

$$p_{t,k} = ((nkw_{-1}\ \beta)/((\Sigma nkw_{-1}\text{-}1)+(V\beta))) * ((ndk_{-1}\ \alpha)/((\Sigma ndk_{-1}\text{-}1)+(K\ \alpha))) \quad (3)$$

III. Research Methodology

The process that occurs in IR underlies the methodology undertaken in this paper as in Fig. 1. This research begins by collecting data for a collection of documents. Then, the documents in the collection were processed in the text processing stage. The results of text processing were selected to obtain features and extracted to obtain feature weights at the feature selection and extraction stage. Documents in collection are clustered based on the weight of the features obtained, resulting in a collection of document clusters in the document clustering stage. And then, the user information's need was processed at the query processing stage. The document retrieval process in the collection that relevant to the query feature is done at document searching stage.

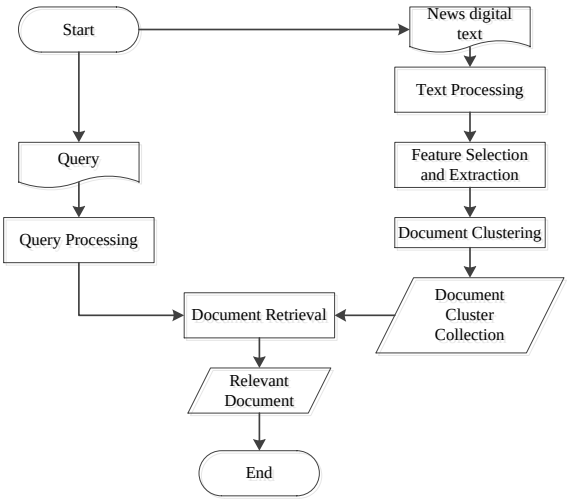

Fig. 1. Research Method

A. Data Source

The collection of documents used in this paper was digital news text taken from the Indonesian news media website. Data was taken from five categories of 125 files; each category consists of 25 files. Data was stored in the form of text files.

B. Text Processing

This stage consisted of four steps: tokenization, case folding, filtering and stemming. Tokenization parsed the text in a news text file into a smaller part, called term or word. Case folding eliminated numbers, symbols and punctuation, and converts the entire character of the term into lowercase. Filtering removed the words listed in the stopwords list; the deleted words are the most common words that appear in the text. Stopwords used consisted of 902 words [14]. Stemming parsed term into prefix, root word, and suffix; deleted prefixes and suffixes, and checked the root word to the basic word dictionary. Rules for removing prefixes and suffixes based on the Nazief-Adriani stemming algorithm, and the basic word dictionary used consisted of 28,527 words [14][15].

C. Feature Selection and Extraction

The feature selection process calculated the term occurrence frequency for each document in the collection based on the TF-IDF concept, and then sorted the TF-IDF results in each document from the largest value. The concept of Luhn's Idea is used to remove terms whose value is at the lower cut-off boundary. Selection features are calculated used the topic-based feature extraction method of Fuzzy Gibbs Latent Dirichlet Allocation (FGLDA). The value of the mixing proportion of documents on the topics and the mixture components of words on the topics were used in this method.

D. Documents Clustering

The process of calculating the distance between documents in the collection used the feature weight value obtained previously using the Cosine Similarity distance function. The distance value obtained is used to group documents using the clustering method of Complete Agglomerative Hierarchy Clustering (CAHC).

E. Query Processing and Document Retrieval

The user information need is written in unstructured text form and processed first with the same process as the text processing of the document collection. The result of text processing in this paper is changed into query with AND logic function.

The document retrieval process was not calculated the distance between the query and the document based on a certain distance function. The retrieval process was done based on a combination of query terms, indexes and cluster collections that have been built before. If a cluster in a collection matched the combination of query terms, the cluster was selected, and all documents in the cluster were displayed as search results. This method was called Document Retrieval based on Topic Clustering method.

IV. Experiment Result

A. Text Processing

The result of document text processing in the collection can be seen in Table 1. In the tokenization results, it can be seen that the text of the document has been parsed into a list of items: '*Angkat*', *'besi* ', *'adalah'*, and so on. From the results of the case folding can be seen that the term *'Angkat* has been changed to lowercase to *'angkat*', the dot symbol on the *'internasional* ' term has been removed. In the filter results it can be seen that the term *'adalah'* has been removed from the result list because the term was in stopword list. In the stemming results it can be seen that the term *'menyumbangkan'* has been replaced with the word *'sumbang*' since the prefix *me-* and suffiks *-kan* have been deleted.

TABLE I. Text Processing Results

Tokenization		**Case Folding**	**Filtering**	**Stemming**
Indonesian	***English***	***Indonesian***	***Indonesian***	***Indonesian***
Angkat	Weightlifting	angkat	angkat	angkat
besi		besi	besi	besi
adalah	is	adalah		
salah	one of the	salah	salah	salah
satu		satu		
cabang		cabang		
olahraga	sport	olahraga	olahraga	olahraga
yang	that	yang		
rutin	routinely	rutin	rutin	rutin
menyum-bangkan	contributes	menyum-bangkan	menyum-bangkan	sumbang
medali	medals	medali	medali	medali
untuk	to	untuk		
Indonesia	Indonesia	indonesia	indonesia	indonesia
di	at	di		
berbagai	various	berbagai		
event	event	event	event	event
interna-sional.	international.	interna-sional	interna-sional	interna-sional

B. Feature Selection and Extraction

The results of the feature selection process can be seen in Table 2 and Table 3. Table 2 showed a list of terms that have been calculated for the TF-IDF weight. For example, the term *'acara*' in document number 1 had an occurrence frequency of 1 time in document with a TF-IDF value of 0.013333. This list of terms was ranked in documents based on the TF-IDF value of the largest value. At Table 3, based on the Luhn's Idea value obtained, the terms that were in Luhn's Idea value highest ranking order was selected, while the rest were removed from the list. So, this new list was a list of features selected by the next process. For example, for document number 1, the value of Luhn's Idea was 9 so that 9 terms were in the lowest ranking order removed, so the term '*acara*', '*alat*', '*anak*', '*apresiasi*', '*asosiasi*' were removed from the list.

The results of feature extraction calculations by the FGLDA method of the selected features are shown in Table 4 and Table 5. For example, Table 4 showed the feature *'baja'* in document number 1 had a weight of 0.5 on the topic number 1. Table 5 showed that each document in the collection was distributed to more than one topic. For example, document number 1 was distributed to topics 1, 3, 4 and 5; document number 2 was distributed to topic number 1, 2, and 3; document number 3 was distributed to all topics. Document weight for each topic had different values. These showed that the FGLDA topic model was able to find the topics contained in a document.

TABLE II. TF-IDF RESULTS

Term		TF	TF-IDF
Indonesian	***English***		
acara	event	1	0.013333
alat	tools	1	0.013333
anak	children	1	0.013333
apresiasi	appreciation	1	0.013333
asosiasi	association	1	0.013333
bahan	material	1	0.013333
baja	steel	8	0.106667
baku	basic	1	0.013333
bangsa	nation	1	0.013333
bangun	build	1	0.013333
banjir	flood	1	0.013333
bijak	wise	2	0.026667
butuh	need	3	0.040000
capai	achieve	1	0.013333
catat	record	1	0.013333

TABLE III. FEATURE SELECTION RESULTS

Term		TF	TF-IDF
Indonesian	***English***		
baja	steel	8	0.106667
bangun	build	1	0.013333
banjir	flood	1	0.013333
bijak	wise	2	0.026667
butuh	need	3	0.040000
capai	achieve	1	0.013333
catat	record	1	0.013333

TABLE IV. FGLDA RESULTS

Term		FGLDA Value	Topic Number
Indonesian	***English***		
baja	steel	0.500000	1
bangun	build	0.818182	1
banjir	flood	0.500000	1
bijak	wise	0.500000	1
butuh	need	0.500000	1
capai	achieve	0.684211	1
catat	record	0.727273	1

TABLE V. DISTRIBUTED TOPIC-DOCUMENT RESULTS

Topic Number	Document Number	TDT Value
1	1	0.681250
2	1	0.000000
3	1	0.118750
4	1	0.093750
5	1	0.106250
1	2	0.754902
2	2	0.107843
3	2	0.137255
4	2	0.000000
5	2	0.000000
1	3	0.533333
2	3	0.104762
3	3	0.123810
4	3	0.114286
5	3	0.123810

C. Document Clustering

The results of calculating the distance between documents in the collection using the FGLDA weights were shown in Table 6. For example, document number 76 had a Cosine Similarity distance to document number 1 of 0.108033, to document number 2 of 0.0, to document number 3 of 0.109833, and so on. Based on the distance between documents in the collection clustering is done with CAHC method and the results as in Table 7. It was seen that documents numbered 81 and 83 were grouped into one cluster in cluster number 129; then the cluster was grouped again with document number 82 on cluster number 131, and so on. For the process of finding documents relevant to the query, each feature was indexed to the corresponding document as shown in Table 8. For example, the feature '*baja*' was indexed to documents number 1 and 2.

TABLE VI. COSINE SIMILARITY RESULTS

Document Number	Document Number	Cosine Value
76	1	0.108033
76	2	0.000000
76	3	0.109833
76	4	0.115629
76	5	0.073533
76	6	0.092757
76	7	0.116939
76	8	0.046831
76	9	0.036149
76	10	0.010106

TABLE VII. CLUSTERING RESULTS

Cluster Number	Document Number
126	86
126	89
127	53
127	54
128	6
128	8
129	81
129	83
130	76
130	80
131	82
131	81
131	83

TABLE VIII. INDEXING RESULTS

Term		Document Number
Indonesian	***English***	
baja	steel	1, 2
bangun	build	1, 2, 3, 21, 23, 24, 69, 72, 111
banjir	flood	1
bijak	wise	1, 12, 14, 15, 25
butuh	need	1, 2
capai	achieve	1, 8, 9, 14, 15, 21, 22, 82, 101, 111
catat	record	1, 6, 8, 9, 10, 78, 107

D. Query Processing and Document Retrieval

In the query processing stage, the user types the information needed in natural language. Examples of information that are typed are as shown in Fig. 2. The text typed by the user is processed through the tokenization, case folding, filtering and stemming as in the processing of document text in collections. The results of query processing can be seen in Table 9.

At the document retrieval stage, each stemming result is matched with a list of indexes that have been formed in the document clustering stage, and the result as in Table 10. These index values become unique features to form queries with AND logic functions as follows.

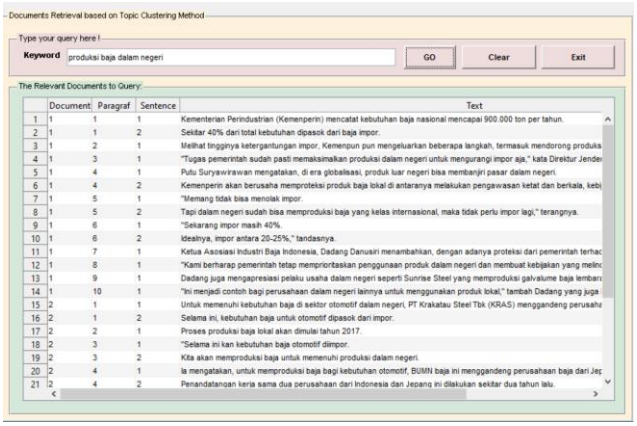

Fig. 2. Documents Retrieval System Interface

TABLE IX. QUERY PROCESSING

Tokenization		Case Folding	Filtering	Stemming
Indonesian	***English***	***Indonesian***	***Indonesian***	***Indonesian***
produksi	production	produksi	produksi	produksi
baja	steel	baja	baja	baja
dalam	in	dalam		
negeri	domestic	negeri	negeri	negeri

TABLE X. QUERY INDEXING

Features		Index of Documents
Indonesian	***English***	***Indonesian***
produksi		1, 2, 18, 20, 31, 33, 34, 35, 59, 62, 103
baja		1, 2
negeri		1, 2, 22, 29, 61, 75, 106, 112

produksi AND *baja* AND *negeri*

(**1, 2**, 18, 20, 31, 33, 34, 35, 59, 62, 103) AND
(**1, 2**) AND
(**1, 2**, 22, 29, 61, 75, 106, 112)

Based on the query above, the document index numbers that match with the query is document number 1 and 2, so that the document retrieval process is carried out on clusters that contain document number 1 and 2 as shown in Fig. 2. It shown that the text in document number 1 and 2 provides information on steel production in Indonesia so that it is relevant to the needs of users of information regarding domestic steel production, which "domestic" refers to the country of Indonesia.

V. CONCLUSION

The improvement of search engine performance has been done in this paper which includes three factors: feature selection, feature extraction and document retrieval. Feature selection has been done using TF-IDF based on Luhn's Idea method. Feature extraction has been done using Fuzzy Gibbs Latent Dirichlet Allocation method. Document retrieval has been done using Document Retrieval based on Topic Clustering method. The results of retrieval process has showed that Document Retrieval based on Topic Clustering method has successfully displayed cluster that have a set of documents that are relevant to queries typed by the user on the search engine. Therefore, this method is able to eliminate the distance calculation process between queries with a set of documents in the collection, so that it is expected to accelerate the retrieval process in the search engine.

ACKNOWLEDGMENT

The authors would like to express the great thank to Directorate General for Research Strengthening and Development, Ministry of Research, Technology and Higher Education, Republic of Indonesia as the sponsor of this research through the Doctoral Dissertation Grant based on Contract of Research No. 013/SP2H/LT/DRPM/2018.

REFERENCES

[1] S. S. Hong, W. Lee, and M. M. Han, "The feature selection method based on Genetic Algorithm for efficient of text clustering and text classification," Int. J. Advance Soft Compu. Appl, vol. 7, 2015.

[2] Z. Masoumeh and K. R. Seeja, "Feature extraction or feature selection for text classification: A case study on phishing email detection," I.J. Information Engineering and Electronic Business, vol. 2, pp. 60-65, 2015.

[3] A. Mahajan and S. R. Sharmistha, "Feature selection for short text classification using Wavelet Packet Transform," in Proceedings of the 19th Conference on Computational Language Learning, Beijing, China, 2015, pp. 321-326.

[4] Y. Wang, Z. Zhou, S. Jin, D. Liu, and M. Lu, "Comparisons and selections of features and classifiers for short text classification," IOP Conference Series: Materials Science and Engineering, vol. 261, 2017.

[5] B. S. Harish and M. B. Revanasiddappa, "A comprehensive survey on various feature selection methods to categorize text documents," International Journal of Computer Applications, vol. 164, 2017.

[6] S. Joseph, C. Mugauri, and S. Sumathy, "Sentiment analysis of feature ranking methods for classification accuracy," IOP Conference Series: Materials Science and Engineering, vol. 263, 2017.

[7] Z. Wu, H. Zhu, G. Li, Z. Cui, H. Huang, J. Li, E. Chen, and G. Xu, "An efficient Wikipedia semantic matching approach to text document classification," Information Sciences, vol. 393, pp. 15-28, 2017.

[8] P. M. Prihatini, I. K. G. D. Putra, I. A. D. Giriantari, and M. Sudarma, "Fuzzy-gibbs latent dirichlet allocation model for feature extraction on Indonesian documents," Contemporary Engineering Sciences, vol. 10, pp. 403-421, 2017.

[9] P. M. Prihatini, I. K. G. D. Putra, I. A. D. Giriantari, and M. Sudarma, "Indonesian text feature extraction using gibbs sampling and mean variational inference latent dirichlet allocation," presented at the Quality of Research (QIR): International Symposium on Electrical and Computer Engineering, 2017 15th International Conference on, Nusa Dua Bali Indonesia, 2017.

[10] C. D. Manning, P. Raghavan, and H. Schutze, An Introduction to Information Retrieval. England: Cambridge University Press, 2008.

[11] I. Kocabas, B. T. Dincer, and B. Karaoglan, "Investigation of Luhn's claim on information retrieval," Turk J Elec Eng & Comp Sci, vol. 19, pp. 993-1004, 2011.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[13] G. Heinrich, Parameter estimation for text analysis: University of Leipzig, Germany, 2008.

[14] P. M. Prihatini, I. K. G. D. Putra, I. A. D. Giriantari, and M. Sudarma, "Stemming algorithm for Indonesian digital news text processing," International Journal of Engineering and Emerging Technology, vol. 2, 2017.

[15] J. Asian, H. E. Williams, and S. M. M. Tahaghoghi, "Stemming Indonesian," presented at the Proceedings of the Twenty-eighth Australasian conference on Computer Science, 2004.