Automated Text Summarization for Indonesian Article Using Vector Space Model

Automated Text Summarization for Indonesian Article Using Vector Space Model

C Slamet[1*], A R Atmadja[1], D S Maylawati[1], R S Lestari[1], W Darmalaksana[2] and M A Ramdhani[1]

[1] Teknik Informatika, Fakultas Sains dan Teknologi, UIN Sunan Gunung Djati, Bandung, Indonesia
[2] Fakultas Ushuludin, UIN Sunan Gunung Djati, Bandung, Indonesia

*cepy_lucky@uinsgd.ac.id

Abstract. In a scientific work, an abstract always contains main information of an article including at least a researched problem, aim(s), methodology, and result of the study. Writing an abstract requires a conscientious analysis since the contents would affect both the readers' interestedness and disinterestedness on a particular or overall research topic. However, people generally write manually by summarizing the article. The aim of this study is constructing automation for summarizing Indonesian articles as an alternative approach to an abstract. This is involving two methods to summarize an article. A Term Frequency-inverse Document Frequency is used to get a keyword and weight terms, and a Vector Space Model is utilized to represent abstract text into a vector that used to identify the linkage of documents. From this method, the result of the summary can be generated from documents. Supporting this research, we used several journal articles written by a manual abstract. The results of this application show that the automatic summarization produces a paragraph which consists of more than three same sentences constantly as compared to manual paragraphing.

1. Introduction

Scientific papers are products of intellectual processes of research, observation, investigation, and long thoughts towards a particular topic or problem. They commonly consist of several parts, one of which is called abstracts which describe the whole part of the paper(s) in brief. Generally, abstracts contain 150 to 250 words covering the aforementioned commonly-written parts. Abstract writing plays an important role since the content of abstracts affect readers' interest of a paper [1].

Text summarization is a process of resuming a text, which means reducing the length of the original text, by using certain words or sentences that representatively contain the main information or description of a particular document. In the meantime, automated text summarization is a summarization process to produce a shortened version of a text using computers aiming at emphasizing the main information or description of a text. This technology is able to help readers acquire the main information of a text through a summary without reading the whole text or document [2]. Automation on document summary has actually been developed since 1958 [2].

Studies in relation to automated text summarization system have been conducted by employing a variety of methods using a variety of languages, some of them are ([2][3][4][5][6]). Those studies utilize a number of algorithms such as Vector Space Model (VSM) and Weighted Tree Similarity (WTS). In a more specific context, it is revealed that VMS method has more recall results than WTS [7]. VSM is commonly used to measure the similarities of texts being tested by calculating the cosine from vector angles made by the document from the keywords vector [2]. Automatically-summarized texts of scientific papers should become abstracts that do not only have similarities in terms of textual context, but also represent the main ideas of the researcher(s). This study is aims to utilize VSM on an application of automated text summarization as an alternative way of making abstracts in Indonesian language and to compare the output of the abstract with those made manually.

2. Literature Review

2.1 Automated Text Summarization

Text summarization, even though it is the short version of a whole text, should really still contain the main substance of the text [2]. Furthermore, a summary is a text re-produced from one of more texts with significantly-proportional information and is no more than 50% of the original text [10]. In the meantime, automation is defined as a technology in which a process or a certain procedure can be implemented without any help from human beings. This without human help process or procedure can be implemented using an instructive program combined with a control system running the instruction [11]. Automated text summarization in this study is an application able to summarize a text and to create an abstract.

2.2 Abstracts

An abstract a summarized form about a certain scientific work and it is usually a separate section of a text. It explains brief explanation of the researchers or writers' ideas to the readers [12] and generally consists of 150 to 250 words [1]. The existence of an abstract will also avoid plagiarism by others. A research study will most likely be protected when only its abstract is displayed and uploaded in the internet [12].

2.3 Text Mining

Text mining is one of data mining parts that is defined as a process of information seeking in which users interact with a pile of documents using analytical tools that actually are components in data mining [14]. Meanwhile, data mining is defined as a complication of processes of mining additional values which are not required manually from the database [19]. What makes text mining different from data mining is that its data are often semi-structured or even unstructured. However, both types of mining usually face the same obstacles such as big amount of data, high dimension, and changing structure. In text mining, it usually appears in the forms of complex and incomplete text structure, unclear and unstandardized meaning, and different language.

2.4 Text Preprocessing

Text preprocessing is a process of preparing a text to become data to analyze in the next step [21]. Initial input in this process is in a form of documents [5] [18] [20]. In this study, text preprocessing contains several steps as follows.

1. Sentence fragmenting
   Sentence fragmenting, which is the first process is text preprocessing, is shortening the content of a certain text into a compilation of sentences.
2. Case folding
   This step is a process of changing all the letters in the document into small capital and deleting all characters that are not the alphabet (a to z).
3. Tokenizing
   This step is a process of selecting input string based on words so that there are only single words.
4. Filtering
   This step is a process of omitting "stop list" words that appear frequently but are not descriptive and not related to a certain theme. In Indonesian language, such unimportant words are usually "di", "oleh", "pada", 'sebuah", "karena", and so on.
5. Stemming
   This step is a process of finding a root of each word so that those words will be in their basic forms (stem). This study uses stemming porter for Indonesian language [17].

2.5 TF/IDF (Term Frequency-Inversed Document Frequency)

In the algorithm of TF/ IDF, a certain formula to calculate the weight (W) of each document towards its keywords is performed [2]. The formula is as follows.

$$W_{dt} = tf_{dt} * IDF_t = tf_{dt} \log \frac{D}{df_t} \quad .....(1)$$

In which:

d	= document	IDF	= Inversed Document Frequency
t	= keywords	TF	= number of search words in a document \
W	= weight of document (-d) to words (–t)	D	= total number of documents
df	= number of documents containing searched words		

After the weight (W) of each document is found, the next process is ordering the quality of each document to find out the similarity level of those documents towards keywords. The example of TF-IDF simple implementation of three documents (D) is as follows.

Keywords (kk) = logistic knowledge
Document 1 (D1) = logistic transaction management
Document 2 (D2) = individual knowledge
Document 3 (D3) = in knowledge management, there is logistic knowledge transfer

Table 1 is the calculation of TF/IDF after text preprocessing of documents 1, 2, and three.

Table 1. Sample Calculation of TF/IDF

Token	tf				df	D/df	IDF	W			
	kk	D1	D2	D3			= Log10(D/df)	kk	D1	D2	D3
Manajemen	0	1	0	1	2	1.5	0.176	0	0.176	0	0.176
Transaksi	0	1	0	0	1	3	0.477	0	0.477	0	0
Logistic	1	1	0	1	2	1.5	0.176	0.176	0.176	0	0.176
Transfer	0	0	0	1	1	3	0.477	0	0	0	0.477
Pengetahuan	1	0	2	2	2	1.5	0.176	0	0	0.176	0.352
individu	0	0	0	1	1	3	0.477	0	0	0.477	0
Total								0.352	0.829	0.653	1.181

From Table 1, it is found that the weight (W) of document 1 (D1) from two words comprising "logistic" and "knowledge is 0.176. However, if there are two documents with the same quality (D1 and D2, for instance), there needs to be an algorithm calculation namely vector-space model. The main idea of this method is by calculating the cosine values of two vectors, which are W from each document and W from each keyword.

2.6 Vector Space Model

Similarities among documents are represented as bag-of-words and are convertible into a vector space model (VSM). In this model, each document and the database and query of the users is represented by a multi-dimension vector. The dimension is in accordance with the term number in the document using this model.

- Vocabulary. It is the rest of different words in the document after preprocessing process that contains t term index. Those terms will then form a vector space;
- Each i term in the document of query j is given weight with real Wij;
- The document and query are expressed as a t vector dimension of dj = (W1, W2, …, Wtj) and there is a certain n document in the collection, which is j = 1, 2, …n.

The example of vector space model of two documents (D1 and D2), one query user (Q1), and three terms (T1, T2, and T3), is shows in Figure 1.

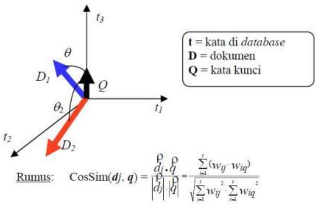
t_3
θ
D_1
Q
θ_2
t_1
t_2
D_2
t = kata di *database*
D = dokumen
Q = kata kunci
Rumus: $\text{CosSim}(\boldsymbol{dj}, \boldsymbol{q}) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \cdot |\vec{q}|} = \frac{\sum_{i=1}^{t}(w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} w_{ij}^2 \cdot \sum_{i=1}^{t} w_{iq}^2}}$

Figure 1. *Vector Space Model* [2]

In vector space model, the collection of documents is represented by term-document matrix or term-frequency matrix. Each cell in the matrix is according to the weight given by a term in a particular document. If the value is zero, it means that the term is not present in the document. The example of term-document matrix for database with t-term document is presented in Figure 2.

$$\left[\begin{array}{ccccc} & T_1 & T_2 & \cdots & T_t \\ D_1 & w_{11} & w_{21} & \cdots & w_{t1} \\ D_2 & w_{12} & w_{22} & \cdots & w_{t2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ D_n & w_{1n} & w_{2n} & \cdots & w_{tn} \end{array}\right]$$

Figure 2. Example of VSM Matrix [2]

Long documents are often considered more relevant in comparison with the short ones. The fact is, it is not always like that. To reduce the influence of the length of the document(s), there is another factor used in the weight namely nominalization of document length. The nominalization used in this case is cosine nominalization displayed below.

$$w(word_i) = \frac{w(word_i)}{\sqrt{w^2(word_1)+w^2(word_2)+\cdots+w^2(word_n)}} \;.....\; (2)$$

In which W = weight from the query and document.

After the cosine values of each document are gained, the results of the weight are put in order. Documents with high weight is prioritized is documents in relation to keywords.

3. Architecture of System

The architecture of the system built involves several components that are related to each other so that this application can finally produce an abstract from an article. Whereas, the data used are articles in the forms of PDF. Those articles are then converted into texts. The texts will then proceed to preprocessing step which comprises sentence fragmenting, case folding, tokenizing, filtering, and stemming. Further, TF-IDF weight process of the text in this study comes up with 20 keywords (features). VSM is then used to compare the similarities of the text and keywords based on the calculation of cosine similarity. Ten sentences as the results of the cosine calculation are the final results namely an abstract. The architecture of the system in automated text summarization of journals can be seen in Figure 3.

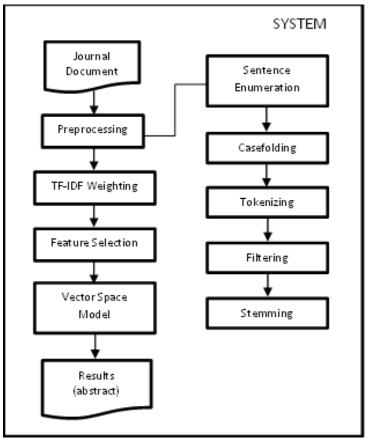

Figure 3. Architecture of System

4. Assessment of System

4.1 Assessment Scenario

Assessment process in this study consists of four steps covering data preparation, preprocessing, TF-IDF weight calculation, keyword similarity selection using VSM which is called cosine similarity.

The data used for the assessment is articles taken from an informatics journal written in Indonesian language. The journal has certain criteria such as omission of header and footer, titles, sub-titles, and tables and figures. The file type of the journal is PDF.

There are five documents tested in this study. The PDF documents are going through parsing process of the text. Table 2 shows the list of the documents tested in the study.

Table 2. Document Test data

#Doc	Doc. Title
D1	Implementation of Vector Space Model and Term Frequency Inverse Document Frequency (TF-IDF) Model to on Information Retrieval System
D2	Facial Recognition System using Template Matching Method
D3	Automated News Summarization based on Text Mining using Generalized Vector Space Model: A Case Study on News Retrieved from Online Mass Media
D4	Classification System of Decorative Plants of Philodendron Leaves using K-Nearest Neighbor (KKN) Method based on Hue, Saturation, and Value (HSV)
D5	Implementation of Ant Colony Optimization Algorithm on Worship Place Searcher Application in Bandung

The next step is preprocessing. This step consists of such processes as sentence fragmenting, case folding, tokenizing, filtering, and stemming. The stemming technique used in this study is stemming by Nazied and Adriani. In this version of stemming, the first action is deleting suffixes such as –lah, -kah, -mu, -nya, and derivational suffixes such as –i, -and, and –kan [17]. To reach the weight of the words, TF-IDF is carried out. The next step is using VSM to check the similarities between the texts and the keywords. The formula used in cosine similarity. The results of this process are a collection of sentences with highest rank of frequency.

4.2 Results of the Test

Table 3. shows the results of the tests between documents with automated abstract and those with manual ones. The comparison uses number 1 if there is a similarity between the automated and manual abstracts.

Table 3. Document Testing

#Doc	System									
	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10
D1	0	1	0	0	1	0	0	1	1	0
D2	0	0	0	0	0	0	0	0	0	0
D3	1	1	0	0	0	0	0	1	0	0
D4	1	1	1	0	1	0	0	0	0	0
D5	0	0	1	1	1	0	0	0	0	0

Table 3. describes that the most similarities lie on documents D1 and D4 with four similarities while the least ones lie on document D2 where there is no similarity at all. Thus, the accuracy level of the system, particularly in making automated abstract, is still low in comparison with the manual ones. This is interpreted based on the similarity level that merely reaches 40% in each document.

5. Conclusion

Vector Space Model (VSM) is used to summarize journals to finally produce abstracts. The results show that the comparison of making abstracts using the system and using manual way comes up with the highest similarities of four sentences. This is due to the fact that sometimes manual abstracts contain words that are not in the body of the texts. Writers usually have their own words in abstracts. In the meantime, automated text summarization consists of several steps such as preprocessing, gaining weight through TF-IDF, and word similarity test using VSM. The results of the summary, which will later become an abstract, is ranked based on the scores on the VSM.

References

[1] B Prasetyo and L M Jannah 2005 *Metode Penelitian Kuantitatif : Teori dan Aplikasi* (Jakarta: PT Raja Grafindo Persada)

[2] Dwijawisnu and A Hetami 2015 *Perancangan Information Retrieval (Ir) Untuk Pencarian Ide Pokok Teks Artikel Berbahasa Inggris Dengan Pembobotan Vector Space Model* **9** 1

[3] N Komang, M Karmila, M Windu and A Kesiman 2013 *Aplikasi Automated Text Summarization (ATS) Pembuat Lead (Teras) Berita dengan Text to Speech (TTS) Menggunakan Algoritma TF-IDF* **2** 807–817

[4] D Nugraha, P Putra and A Suharsono 2008 *Peringkas Dokumen Tunggal Berbahasa Indonesia Menggunakan Metode Sentences Clustering dan Frequent Term* 1–10

[5] M Mustaqhfiri and Z Abidin 2001 *Peringkasan Teks Otomatis Berita Berbahasa Indonesia Menggunakan Metode Maximum Marginal Relevance*

[6] A Yuliawati, D Purwitasari and U L Yuhana 2011 *Implementasi peringkasan otomatis pada dokumen terstruktur dengan metode faktorisasi matriks nonnegatif* 1–9

[7] V B Wicaksono and S W Sihwi 2016 *Analisis Perbandingan Metode Vector Space Model dan Weighted Tree Similarity dengan Cosine Similarity pada kasus Pencarian Informasi Pedoman Pengobatan Dasar di Puskesmas* 1–11

[8] S R A and M Shalahuddin 2005 *Rekayasa Perangkat Lunak (Terstruktur dan Berorientasi Objek)* Bandung

[9] R A and M Shalahuddin 2015 *Rekayasa Perangkat Lunak*

[10] M N Rachmatullah 2015 *Implementasi Jaringan Syaraf Tiruan Pada Sistem Peringkasan* Sentika

[11] Nengsih 2006 *Indonesia Menggunakan Algoritma Term Frequency Dan Lead Indonesian Language Automatic Summarizer for Single*

[12] Rachmandianto, “ABSTRAK,” 2011.

[13] E Prasetyo 2012 *Data Mining Konsep dan Aplikasi menggunakan MATLAB* Yogyakarta

[14] J Han and M Kamber 2006 *Data Mining: Concepts and Techniques* San Francisco: Diane Cerra

[15] W Hastomo 2013 *Pengertian dan Kelebihan Database MySQL*

[16] A Saputra *Smarty PHP OOP Engine for PHP Template*

[17] Adriani M, Asian J, Nazief B, Tahaghoghi S M and Williams H E 2007 Stemming Indonesian: A confix stripping approach *ACM Transactions on Asian Language Information Processing (TALIP)* **6** 4 1-33

[18] Atmadja A R and Purwarianti A 2015 Comparison on the rule based method and statistical based method on emotion classification for Indonesian Twitter text. *In Information Technology Systems and Innovation (ICITSI)* 2015 International Conference on (pp. 1-6) IEEE

[19] Sandi G, Supangkat S H and Slamet C 2016 Health Risk Prediction for Treatment of Hypertension *Cyber and IT Service Management, International Conference*. 1-6 IEEE

[20] Slamet C 2016 Clustering the Verses of The Holy Qur’an Using K-Means Algoritm *Asian Journal of Information Technology* 5159-5162 Medwell Journals.

[21] Maylawati D S and Saptawati G A P 2016 Set of Frequent Word Item sets as Feature Representation for Text with Indonesian Slang *International Conference on Computing and Applied Informatics* 1-6 IOP Publishing