Searching Information Tourism using Vector Space Model

Dede Wintana[1], Sfenrianto[2], Hikmatulloh[1], Mugi Raharjo[1], Jordy Lesmana Putra[1], Dyah Ayu Ambarsari[1], Desi Dwi Jayanti[1]

[1]*Master of Computer Science Postgraduate Program, STMIK Nusa Mandiri, Jakarta, Indonesia*

[2]*Information System management Departemen, BINUS Graduate Program, Master of Information System Management, Bina Nusantara University, Jakarta, Indonesia*

Keywords: Tourism, Searching, Vector Space Model, Waterfall, Information Retrival.

Abstract: Development of tourism is directed at creating tourist destinations that are evenly distributed in an area, including tourism in Sukabumi, West Java, Indonesia. This study conducts secondary data retrieval from the internet by using keyword "air terjun sukabumi". Information search method is to use a Vector Space Model (VSM) to see the level of similarity with the weighting term. By passing various processes such as tokensizing, filtering, stemming, tf and df, calculation of inverse document frequency and several other processes. The results shows that the proposed algorithm is very suitable for processing the sample. Determination of document ranking is done by several stages of the algorithm that produce some documents with the highest value. The study can also prove the proposed algorithm can work well in the case of searching for waterfall data in Sukabumi district.

1 INTRODUCTION

West Java has one district with promising potential for the advancement of tourism, namely Sukabumi Regency, which is located in the Java south. In the Regional Long Term Development Plan (RPJPD) of Sukabumi Regency in 2005-2025 it was stated that the priority of tourism development was directed at the creation of tourism destinations in Sukabumi as one of the leading tourism destinations in West Java. Indonesian tourism competition was increasingly sharp, so that demanding every region explore potential sources power to sell, attract and be visited by tourists (Darsiharjo, 2016). Eco-tourism is a type or type of tourism that makes natural resources an object that is one of the selling power of Sukabumi district, coupled with artificial resources. For some areas in West Java, eco-agro-tourism has developed well, but there are still many other areas that have the potential to develop eco-agro-tourism for the advancement of the region and the welfare of its people (Hasugian, 2003).

From a development perspective, ecotourism businesses should only be considered 'successful' if local communities have some control over them and if they share the benefits equitably emerge from ecotourism activities (Scheyvens, 1999). One of the visited areas at Sukabumi is the existing Ciletuh geopark in the southern region of Sukabumi. The Geopark is an earth park that is included in a conservation area, which has elements of geodiversity, biodiversity, and cultural diversity which has aspects in the field of education as knowledge geology on the uniqueness and diversity of the earth's heritage in managing the area as tourism (Darsiharjo, 2016). It has also many diversity of tourist attractions such as beaches and many waterfalls that can be visited and waterfalls in the Ciletuh area (Hardiyono et al., 2015). There are some waterfalls in Ciletuh, namely: Awang waterfall, Curung Puncak Manik, Sodong waterfall, Cimarinjung waterfall and much more.

Collaboration is very important for tourism marketing that is sucses, purpose, and electronic communication are a new opportunity to be an opportunity to work between tourism suppliers (Palmer and Mccole, 2009). The increasing number of tourists to Sukabumi is one way to increase regional income, but until now tourism development is still not evenly distributed, especially the many waterfalls, consequently even distribution of tourism objects is not yet optimal so scientific studies are needed on tourist attraction in Sukabumi.

Many countries promote nature-based tourism in Indonesia to promote the purpose of nature conservation and income generation (Hearne and Salinas,

Wintana, D., Sfenrianto, ., Hikmatulloh, ., Raharjo, M., Putra, J., Ambarsari, D. and Jayanti, D.
Searching Information Tourism using Vector Space Model.
DOI: 10.5220/0009882102470251
In *Proceedings of the 2nd International Conference on Applied Science, Engineering and Social Sciences (ICASESS 2019)*, pages 247-251
ISBN: 978-989-758-452-7

2002). For natural resource management planning planners make schemes and approaches to anticipate various biased information caused by these limitations. The characteristics of coastal resources and Ciletuh geopark are quite complex and require specific management policies in order to provide optimal and sustainable life and livelihood (Wahyudin, 2011).

In this study conducted research from the Internet by using the keyword "air terjun sukabumi" to determine the extent to which tourists know about tourist objects in the region. It is expected that with using a Vector Space Model (VSM) can be known about which regions little tourists for help the Sukabumi tourism department in promoting tourism objects.

2 RESEARCH FRAMEWORK

The research framework starts from data presentation (see Figure 1). The data that will be used is secondary data taken from the internet related waterfall in Sukabumi on the internet. The data will be processed using vector space model (SVM) to calculate the amount of value obtained based on a keyword.

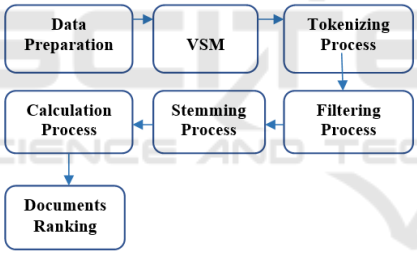

Figure 1: Research Framework.

The VSM is the basic method used for text representation. To represent the text of each feature term (Ti) considered as a coordinate in n-dimensional vector space, the corresponding weight (W1, W2, W3, ... Wn) is considered to be the coordinate of value. Then, it can be used to represent text (Houy, 2013) (Langcai et al., 2017). It is a method to see the level of proximity or similarity term by weighting the term. Documents are viewed as a vector that has a magnitude (distance) and direction. In the VSM, a term is represented by a dimension of vector space. The relevance of a document to a query is based on the similarity between the document vector and the query vector (Amin, 2013) stages in the Information Retrieval covering several processes in documents.

Then, indexing process to get the weight of each term in the document. Calculation of these weights is done by calculating the Term Frequency (Tf) and Document Frequency (Df) of each term contained in the document collection. Next process are tokenizing, filtering, and stemming.

The tokenizing process is done by a mechanism if the document on the corpus is found in a space, so the term between spaces will be retrieved by the system. Then the term is placed in the initial table. Process results in the form of original terms (terms that still have additions, inherent punctuation marks, and numbers).

The filtering process is done by a mechanism if the terms in the initial table table are found punctuation, capital letters, and numbers. Then the program will remove (punctuation and numbers) and change capital letters to lowercase letters, then check the term with stopwords. The process results in the form of selected terms (without punctuation, without capital letters, and not including stopwords).

The Stemming process is a program for eliminating process or how to remove the rewards found in the filtering term. The eliminating process is done by removing the prefix, insertion, and suffix. The result of the process is frekuency table. Finally the calculation process and ranking of documents are carried out.

3 RESULT AND DISCUSSION

3.1 Data Preparation

The research data was obtained from searches on Google's search engine by inputting the request "air terjun sukabumi". There are 8 (eight) document sample titles (D1 - D8) search results, including:

[D1] Curug Sawer, Air Terjun Eksotis di Sukabumi

[D2] Air Terjun Cikaso, Wisata Alam Andalan Kabupaten Sukabumi

[D3] Air Terjun Pareang, kemegahan surga tersembunyi di Sukabumi

[D4] Mengintip Blue Curug Cikaso yang mempesona di Sukabumi Selatan

[D5] Curug Mawi, tempat baru untuk berada di Cibadak Sukabumi

[D6] Air Terjun Gerong Sukabumi Air Terjun Instagramable

[D7] Eksotisme Air Terjun Cimarinjung, Wisata Dunia Jurrasic dalam gaya Sukabumi

[D8] Curug Caweni, Alam Perawan Sukabumi

3.2 Tokenizing

After data is available, the next step is tokenize process. Figure 2 shows the results of the tokenize which have to group each word from the document that has been obtained.

D1	D2	D3	D4
Curug	Air	Air	Mengintip
Sawer	Terjun	Terjun	Blue
Air	Cikaso	Pareang	Curug
Terjun	Wisata	kemegahan	Cikaso
Eksotis	Alam	Surga	yang
di	Andalan	tersembunyi	mempesona
Sukabumi	Kabupaten Sukabumi	di Sukabumi	di Sukabumi Selatan
D5	**D6**	**D7**	**D8**
Curuk	Air	Eksotisme	Curug
Mawi	Terjun	Air	Caweni
tempat	Gerong	Terjun	Alam
baru	Sukabumi	Cimarinjung	Perawan
Berada di	Air Terjun	Wisata Dunia	Sukabumi
Cibadak	Instagramable	Jurrasic	
Sukabumi		dalam gaya Sukabumi	

Figure 2: Tokenizing Process.

3.3 Filtering

After tokenizing, the words that appear in the data are known. It has been applied for a few words, then it will enter the filtering stage.

Figure 3 - Figure 4 shows the filtering results from the data.

Air	ala	Alam
Biru	Caweni	Cibadak
Curug	di	Eksotis
Instagramable	Jurrasic	Kabupaten
Mawi	Mengintip	ngadem
Pesona	Sawer	Selatan
Terjun	Tersembunyi	Wisata

Figure 3: Filtering Process.

Andalan	baru
Cikaso	Cimarinjung
Eksotisme	Gerong
Kemegahan	lokasi
Pareang	Perawan
Sukabumi	Surga
World	

Figure 4: Filtering Process (Extension).

3.4 Stemming

In the Filtering process is data that generates 34 words. While in this process of stemming is 31 words (see Figure 5 - Figure 6).

air	alam	andalan
Caweni	cibadak	Cikaso
Eksotis	Gerong	instagramable
kemegahan	lokasi	Mawi
Pareang	perawan	Pesona
Sukabumi	Surge	terjun
World		

Figure 5: Stemming Process.

baru	biru
cimarinjung	curug
jurrasic	kabupaten
mengintip	ngadem
sawer	selatan
tersembunyi	wisata

Figure 6: Stemming Process (extension).

3.5 Calculation

The calculation process consists of tf-idf, document weight (W), distance Q - D, the document calculation of DOT, and similarity. In the VSM approach, calculations can be found based on the term frequency

and inverse document frequency $(tf - idf)$ based on the log Number of Documents (D) / many documents (df). For example "Air" data, It is known: D = 8 ([D1], [D2], [D3], [D4], [D5], [D6], [D7], and [D 8]); df (water) = 2; then $tf - idf = log(8/2) = 0.6021$. Then to calculate document weights using equations $wf(td) = tf(td) * idft$. Based on document "air" data in document 1 [D1], then $tf(td) =; idf = 0.602. Thus wf(td) = 1 * 0.6021 = 0.6021$.

To calculate the distance query (Q) to document distance (D) using the equation : $Sqrt(\sum_{j=1}^{n} Q_j^2)$. DOT product calculations use the equation: $\sum(Q * D_i) = \sum_{=1}^{n} Q_j D_{ij}$. The result of $Q - D$ and DOT Product from documents can be seen in Figure 7 and Figure 8.

Token	Q-D								
	Q^2	D1^2	D2^2	D3^2	D4^2	D5^2	D6^2	D7^2	D8^2
air	0	0,36248	0	0	0	0	0,36248	0	0
alam	0	0	0,36248	0	0	0	0	0	0,36248
andalan	0	0	0,81557	0	0	0	0	0	0
baru	0	0	0	0	0	0,81557	0	0	0
biru	0	0	0	0	0,81557	0	0	0	0
caweni	0	0	0	0	0	0	0	0	0,81557
cibadak	0	0	0	0	0	0,81557	0	0	0
cikaso	0	0	0,36248	0	0,36248	0	0	0	0
cimarinjung	0	0	0	0	0	0	0	0,81557	0
curug	0	0	0	0	0	0	0	0	0
eksotis	0	0,81557	0	0	0	0	0	0	0
gerong	0	0	0	0	0	0	0,81557	0	0
instagramable	0	0	0	0	0	0	0,81557	0	0
jurrasic	0	0	0	0	0	0	0	0,81557	0
kabupaten	0,81557	0	0,81557	0	0	0	0	0	0
kemegahan	0	0	0	0,81557	0	0	0	0	0
lokasi	0,81557	0	0	0	0	0,81557	0	0	0
mawi	0	0	0	0	0	0,81557	0	0	0
mengintip	0	0	0	0	0,81557	0	0	0	0
ngadem	0	0	0	0	0	0,81557	0	0	0
pareang	0	0	0	0,81557	0	0	0	0	0
perawan	0	0	0	0	0	0	0	0	0,81557
pesona	0	0	0	0	0,81557	0	0	0	0
sawer	0	0,81557	0	0	0	0	0	0	0
selatan	0	0	0	0	0,81557	0	0	0	0
sukabumi	0,36248	0,36248	0	0	0	0	0,36248	0	0
surga	0	0	0	0,81557	0	0	0	0	0
terjun	0,36248	0,36248	0	0	0	0	0,36248	0	0
tersembunyi	0	0	0	0,81557	0	0	0	0	0
wisata	0	0	0,36248	0	0	0	0	0,36248	0
world	0	0	0	0	0	0	0	0,81557	0

Figure 7: Document Calculation of Distance.

DOT							
Q*D1	Q*D2	Q*D3	Q*D4	Q*D5	Q*D6	Q*D7	Q*D8
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	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	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0,665157	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0,665157	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	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0,131389	0	0	0	0	0,131389	0	0
0	0	0	0	0	0	0	0
0,131389	0	0	0	0	0,131389	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

Figure 8: Document Calculation of DOT product.

The next process is to calculate the cosine angle between the vector Query with each document by the formula: Consine θ $D_i = (Q * D)/(|Q| * |D_i|)$. The results of each document are Cosinus [D1] = ((1.53) * (1,648809)) / (0,262778); Cosinus [D2] = ((1.53) * (1,648809)) / (0,665157); Cosinus [D3] = ((1.53) * (1,80618)) / (0); Cosinus [D4] = ((1.53) * (1,903881)) / (0); Cosinus [D5] = ((1.53) * (2,019371)) / (0,665157); Cosinus [D6] = ((1.53) * (1,648809)) / (0,262778); Cosinus [D7] = ((1.53) * (1.676064)) / (0); and Cosinus [D8] = ((1.53) * (1,411956)) / (0).

3.6 Document Ranking

From the results of the Vector Space Model (VSM) analysis, the ranking results for the top four of the 8 documents that have been calculated are Ranking 1: [D2] with a value of 0.262819; Ranking 2: [D5] with a value of 0.21459; Ranking 3: [D1] with a value of 0.10383; and Ranking 4: [D.6] with a value of 0.10382. So the documents that are most relevant to the keywords "air terjun sukabumi" is the document 4 or D4 = Mengintip Blue Curug Cikaso yang mempesona di Sukabumi Selatan.

4 CONCLUSION

The study results show that the proposed algorithm can function quite well in processing a sample of 8 (eight) documents. Determination of document ranking is done by several stages of the VSM (Vector Space Model) algorithm that produces 4 (four) large documents with the highest value, and the value 0.262819 becomes the highest value. It can prove that the proposed algorithm can work well in the case of search for keyword "air terjun sukabumi".

The development of this research still has to be done, because in this study the data taken is still in the scope of the Sukabumi district, so that more optimized results can then be taken in further development within the scope of a wider waterfall data. In further research can also be developed by trying several algorithms that are more optimal in this study case 6.

REFERENCES

Amin, F. (2013). Sistem temu kembali informasi dengan pemeringkatan metode vector space model. 2(2):122–129.

Darsiharjo, D. (2016). Pengembangan geopark ciletuh berbasis partisipasi masyarakat sebagai kawasan geowisata di kabupaten sukabumi. 13(1).

Hardiyono, A., Syafri, I., Rosana, M. F., Yuningsih, E. Y., and Andriany, S. S. (2015). Potensi geowisata di kawasan teluk ciletuh, sukabumi, jawa barat. 13(2).

Hasugian, J. (2003). Penggunaan bahasa alamiah dan kosa kata terkontrol dalam sistem temu kembali informasi berbasis teks.

Hearne, R. R. and Salinas, Z. M. (2002). The use of choice experiments in the analysis of tourist preferences for ecotourism development in costa rica. 64(2):153–163.

Houy, L. M. (2013). A generalized framework for ontology-based information retrieval. pages 165–169.

Langcai, C., Zhihui, L., and Yuanfang, L. (2017). Research of text clustering based on improved vsm by tf under the framework of mahout. pages 6597–6600.

Palmer, A. and Mccole, P. (2009). International journal of contemporary hospitality management emerald article : The role of electronic commerce in creating virtual tourism destination marketing organisations the role of electronic commerce in creating virtual tourism destination marketing. 12:198–204.

Scheyvens, R. (1999). Ecotourism and the empowerment of local communities. 20:245–249.

Wahyudin, Y. (2011). Karakteristik sumberdaya pesisir dan laut kawasan teluk palabuhanratu , kabupaten sukabumi , jawa barat. 1:19–32.