OPEN ACCESS
Indo-JC
Indonesia Journal of Computing
ISSN 2460-9056
socj.telkomuniversity.ac.id/indojc

Ind. Journal on Computing
Vol. 5, Issue. 1, March 2020. pp. 103-114
doi:10.21108/indojc.2020.5.1.397

Ontology-Based Semantic Search on Tourism Information Search System

Ranestari Sastriani [#1], Z K Abdurahman Baizal [*2], Dana Sulistyo Kusumo [#3]

 # School of Computing, Telkom University
Jl. Telekomunikasi No. 01, Terusan Buah Batu, Bandung 40257

[1] ranestari@students.telkomuniversity.ac.id, [2]baizal@telkomuniversity.ac.id,
[3] danakusumo@telkomuniversity.ac.id

Abstract

Nowadays, search engines are used to find various information on the Web. Search engines that are implied traditional search based on keywords do not always provide relevant information. Meanwhile by semantic search can provide relevant results because it can understand the meaning of the context and its connection. One approach to semantic search is to use ontology. Ontology is used to define concepts and relevant relations in a domain. In *Information Retrieval* (IR), an ontology encompasses in returning relevant information from a collection of unstructured information. One application of ontology in IR is query expansion. In this research, The query entered by the users will be searched for its relation to the ontology domain which will be expanded later. We use tourism in Bandung Raya as an ontology domain. The search in this research is represented by using Vector Space Model (VSM). Furthermore, relevant documents from search result will be provided to users. Based on the results of testing, the system that is built is able to provide relevant information with a higher MAP value of 0.563. While the search system that only applies the VSM method has a MAP value of 0,315.

Keywords: Semantic Web, Ontology, Semantic Search, Information Retrieval, Vector Space Model

Abstrak

Saat ini, mesin pencari digunakan untuk mencari berbagai informasi di Web. Mesin pencari yang menerapkan pencarian secara tradisional berdasarkan kata kunci tidak selalu memberikan informasi yang relevan. Sedangkan pencarian semantik dapat memberikan hasil pencarian yang relevan karena dapat memahami makna konteks dan keterhubungannya. Salah satu pendekatan untuk pencarian semantik adalah dengan menggunakan ontologi. Ontologi digunakan untuk mendefinisikan konsep dan hubungan yang relevan dalam sebuah domain. Pada *Information Retrieval* (IR), ontologi memiliki peran dalam mengembalikan informasi yang relevan dari berbagai macam infromasi. Salah satu penerapan ontologi dalam IR adalah perluasan kueri. Pada penelitian ini, kueri yang dimasukkan pengguna akan dicari keterhubungannya pada domain ontologi untuk diperluas. Domain ontologi yang digunakan adalah pariwisata di Bandung Raya**.** Proses Pencarian pada penelitian ini, direpresentasikan dengan menggunakan Vector Space Model (VSM). Selanjutnya dokumen yang relevan dari hasil pencarian akan diberikan kepada pengguna. Berdasarkan hasil pengujian, sistem yang dibangun mampu memberikan informasi yang relevan dengan nilai MAP yang lebih tinggi yaitu 0,563. Sedangkan sistem pencarian yang hanya menerapkan metode VSM memiliki nilai MAP 0,315.

Kata Kunci: Web Semantik, Ontologi, Pencarian Semantik, Information Retrieval, Vector Space Model

Received on Jan 2020. Accepted on March 2020

I. Introduction

WEB has become one of the sources of tourist information nowadays. There are a lot of tourist attractions in Indonesia that are interested to be visited by people. The information about tourist attractions is needed for those who want to make a visit. By using search engines, the visitors or the tourist can find information about tourist attractions that they want to visit.

Search engine is one of the systems used for Information Retrieval (IR). A good IR system is able to provide relevant results to users[1]. In search engine, users put keywords that are related to what they are looking for. However, searches based on keywords or commonly called as traditional search do not always provide relevant information to the users because the searches do not pay attention to the meaning of the keywords, but only based on the appearance of the word in the document[2]. This can make the information search process ineffective. However, it does not rule out the possibility that users also see some irrelevant documents if they think the information is useful[3].

Vector Space Model (VSM) is one of the traditional IR techniques. In VSM, documents and queries are represented as vectors[4]. Each term that is appeared will be given a weight using the *term frequency-inverse document frequency* (tf-idf). This weight represents the importance of a word in the user's document and query. The relevance between document and query can be calculated as the angular cosine between two vectors using the calculation of cosine similarity.

The semantic web allows computers to understand the meaning of the information and the relation so that they can be understood by humans or machines. One approach to use the semantic web is using ontology. Ontology is used to define concepts, relations, and other differences that are relevant to domain modeling[5]. In Information Retrieval (IR), ontology has a role in returning relevant information from a collection of unstructured information so that it can improve the accuracy of the search results[2].

Waseem Alroimama [1] built an Arabic retrieval information system with ontology-based query expansion. The ontology domain of noun in Arabic is used for query expansion on the system. The VSM method is used to represent the search index. The system processes the query expansion by searching for the synonyms and relations from the previous query concept. The system is able to improve precision and recall to outperform the results of keyword-based system performance.

In this research, an ontology-based semantic search system in the tourism domain is built to obtaining relevant information. In this system, a traditional information retrieval technique, which is vector space model and expansion of queries using the tourism domain ontology is combined. We use metropolitan area surrounding the city of Bandung, West Java, Indonesia commonly called Bandung Raya as an ontology domain. For the evaluation, calculation of Mean Average Precision (MAP) is used for the evaluation of system built.

II. Literature Review

A. *Information Retrieval*

Information Retrieval (IR) system allows users to find information in a collection of documents or other information sources. The users enter queries to the IR system to retrieve information that might be relevant to users[6]. The IR system consists of the following processes[7]:

1. Indexing, in this process, a document is represented in an inverted index, then the weight value of each word in the document is calculated.
2. Filtering, in this process, all stopword and common words are deleted.
3. Searching is the main process of IR system. Various techniques can be used to retrieve documents that match the user's needs.

The traditional IR model is divided into three categories: algebraic models, theoretic models, and probabilistic models [4]. Hai Dong [4] conducted a survey and analysis of the advantages and disadvantages of traditional

information retrieval models. One of the is the algebraic method VSM (Vector Space Model). From the survey results, the VSM method with weighted tf-idf has advantages in improving the performance of document retrieval processes. However, by increasing documents dynamically can extend the indexing process time. In addition, the VSM method does not pay attention to the relations between terms in the document.

B. *Semantic Search*

Semantic search is a series of techniques used to extract information from structured data like ontology. It also referred to a combinator or evolutionary version of a traditional information search algorithm [4]. Semantic search can increase effectiveness in the Information Retrieval (IR) process because of its ability to provide organized and relevant data using synonyms/meanings of a concept[8]. In general, the process in semantic search is[9]: (1) the extraction of relevant concepts to the user's sentence, (2) the concept seeks for its connection to ontology, (3) the search results are given to the users.

C. *Vector Space Model*

In the Vector Space Model (VSM), document and query are represented as vectors. Each term is given a weight that represents the importance of the term in the document and query. Cosine similarity is one of the methods in VSM. Cosine is used to measure the similarity between two vectors and can be used to rank documents by finding the closest user's query[10]. The cosine similarity formula is shown in equation(1).

$$Sim(d_j, q) \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i}^{t} w^2{}_{i,j}} \times \sqrt{\sum_{i}^{t} w^2{}_{i,q}}} \quad (1)$$

Sim(d_j,q) shows the similarity between document j and query q. $w_{i,j}$ is the weight of term i in document j, and $w_{i,q}$ is the weight of term i in query q. Before calculating the similarity between document and query, we need to calculate the weight of each term in the document and query. *Term frequency-inverse document frequency* (tf-idf) is one of the methods to determine the weight of a term. The calculations of tf-idf are shown in equations (2),(3), and (4) [1][3].

$$tf_{i,j} = 0.5 + 0.5 \left(\frac{tf_{i,j}}{max\ tf_{i,j}}\right) \quad (2)$$

$tf_{i,j}$ is the number of occurrences of term i in document j. $max\ tf_{i,j}$ is the number of terms in document j.

$$idf_i = log \frac{N}{df_i} \quad (3)$$

with idf_i is inverse document frequency, N is the number of documents in the system, and df_i is the number of documents with the occurrence of term i.

$$w_{ij} = tf_{i,j} \times idf_i \quad (4)$$

w_{ij} shows the weight of term i in the document j.

D. *Ontology*

Ontology is used to improve recall and precision in IR processes[6]. Thomas Gurber defined ontology is a specification of a conceptualization [5]. A specification is a form of definition from vocabulary representation (class, relation, etc.). Conceptualization contains concepts of objects and other entities that are assumed to exist in the domain and limits of the use[11]. Ontology is represented by using the language OWL (Ontology Web Language) and RDF (Resource Description Framework)[12].

a) RDF/OWL

RDF is a format data model of an object ('*resource*') in the form of a graph, which is used by the semantic web to store data [5]. Each document of RDF is a collection of statements consisting of subject, predicate, and object, commonly known as triples. Resource Description Framework Schema (RDFS) provides basic vocabulary to describe properties and classes from RDF resources.

OWL (Ontology Web Language) is a semantic web language designed to represent complex information of an entity, group of an entity, and the relationship between various entities[13]. OWL extends RDFS by adding more detailed constructions to describe the semantics of RDF statements [1].

b) SPARQL

SPARQL (SPARQL Protocol and RDF Query Language) is query language specifically designed to RDF query database.

c) Protégé

Protégé was developed by Stanford Center for Biomedical Informatics Research at Stanford University School of Medicine. Protégé is used to create an ontology domain, adjust the form for data entry, and for data entry. Protégé supports various storage formats such as OWL, RDF, XML, and HTML[14].

E. *Evaluation of Information Retrieval System*

During the information searching, it is needed to be translated into a query. Relevant information is valued relative to the information needed, not the query [3]. Recall and precision are one of the tests in the retrieval system. Precision can be considered as the measurement of accuracy, while recall as perfection. In precision calculations, it is necessary to pay attention to the number of relevant documents retrieved, compared to the total number of documents retrieved. While in the recall calculations, it is necessary to pay attention to the relevant documents retrieved compared to the number of relevant documents[15].

A lot of information from various sources makes the quality of ordered search results very important[16]. In recent years, the calculation of Mean Average Precision (MAP) has become more commonly used [3]. MAP is the average precision value obtained from the top of document k every time a relevant document is obtained. The MAP calculation can be shown in followed equation (5).

$$MAP = \frac{\sum_{q=1}^{Q} AP(q)}{Q} \tag{5}$$

Q is the number of queries in the set and AP(q) is the average precision for a given query q.

III. RESEARCH METHOD

The architecture of the system has shown in Fig. 1.

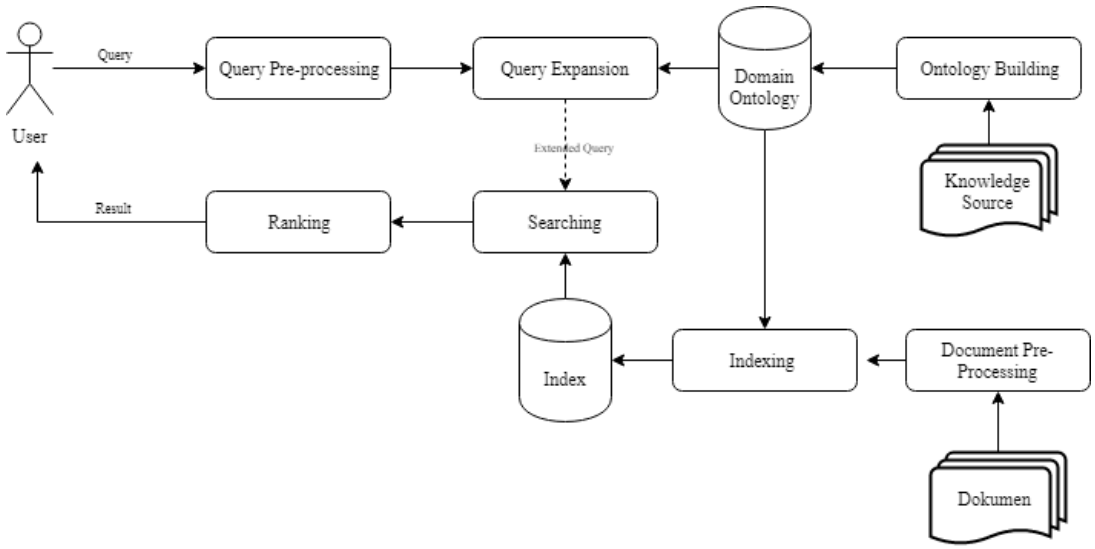


Fig. 1 System Architecture

A. *Ontology Domain*

Ontology is used to expand user queries and indexing process. We make additional individuals and classes related to tourism objects in Bandung Raya. It done manually using Protégé. The ontology domain used has two main classes, penginapan and Tempat_Wisata which have six main subclasses: Alam, Budaya, Kuliner, Olah Raga, Rekreasi, and Rute. Individuals represent places in the Greater Bandung area. Tourism ontology is represented by the Web Ontology Language (OWL).

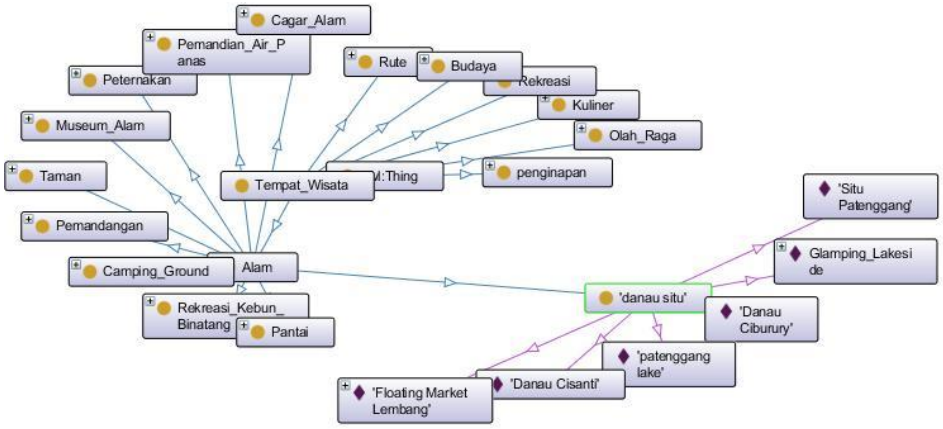


Fig. 2 Example of Class and Individual Graph on Ontolgy Domain

B. *Documents Pre-Processing*

The step in of pre-processing documents has shown in Fig. 3

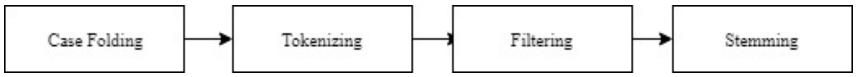


Fig. 3 Pre-processing Document Process

The document used is the documents in the file formats of .txt. We use 237 review articles about tourism in Bandung Raya as document collection[17]. The steps in pre-process of this document are case folding, tokenizing, stemming, and stopword removal. Case folding is the process of converting all letters in a document to lowercase (*Wisata Bandung* become *wisata bandung*) so that all the words in the document will be consistent. The next step is tokenizing process. It is the stage of cutting the input string based on each word that composes it; for example *'Wisata Bandung'* become '*Wisata' 'Bandung'*. After that, there is stopword removal or filtering. At this stage, words like '*dari'*, '*yang'*, '*adalah'*, and other words with high frequency are omitted. The last step is stemming process. S*temming* is the stage of changing words into basic words, such as '*berwisata'* become '*wisata'*. The example of pre-processing is shown in Fig. 4.

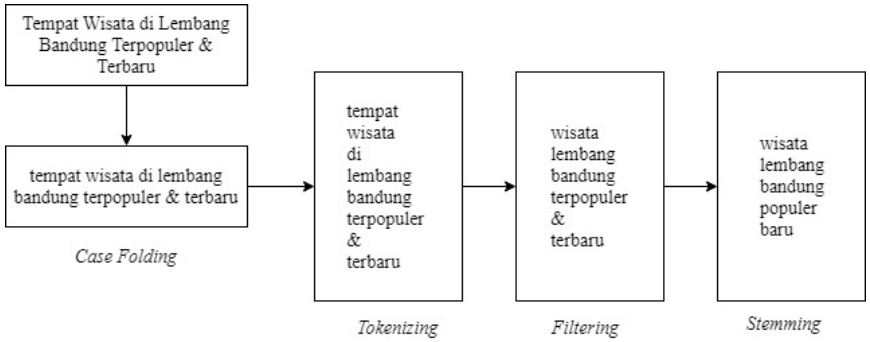
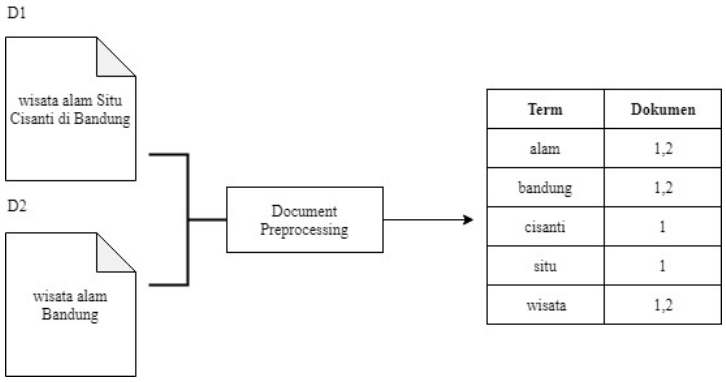


Fig. 4 Example of Pre-processing Document

C. *Indexing Process*

In the indexing process, an inverted index and weighting are made. It is shown in Fig. 5 and Table I. The inverted *index* represents the word that is showing in both document and ontology. Next will be the calculation of the weight of each word in the document using tf-idf weighting calculation by applying the formula in equation (2), (3) and (4).



Term	Dokumen
alam	1,2
bandung	1,2
cisanti	1
situ	1
wisata	1,2

Fig. 5 Inverted Index

TABLE I
TFIDF WEIGHTING

Term	tf		idf	W=tf*idf	
	d1	d2		d1	d2
alam	0.6	0.666	0	0	0
bandung	0.6	0.666	0	0	0
cisanti	0.6	0.5	0.301	0.1806	0.1505
situ	0.6	0.5	0.301	0.1806	0.1505
wisata	0.6	0.666	0	0	0

D. *Query Pre-Processing*

The steps in query preprocessing are the same as the steps in document preprocessing, they are case folding, tokenizing, filtering, and stemming.

E. *Query Expansion*

In this step, each term in the query that matches the concept on the ontology will be expanded by adding all the relations of the concept to the previous query. The relation of each concept is taken using SPARQL language. As an example: if the term of '*danau'* is in the user's query, then the previous query will be expanded with any relations related to the concept of '*danau'* on ontology, as shown in Table II. Based on the tourism ontology, the query will be expanded to '*danau alam situ Floating Market Lembang Glamping Lakeside Situ Cisanti Situ Cuburuy Situ Lembang Situ Patenggang'*.

TABLE II
EXAMPLE OF CONCEPT RELATION

Konsep	Relasi	Konsep
Danau	SubClassOf	Alam
	label	situ
	Instance	Floating Market Lembang
		Glamping Lakeside
		Situ Cisanti
		Situ Ciburuy
		Situ Lembang
		Situ Patenggang

F. *Searching, Document Retrieval, and Document Ranking*

The process of searching is represented using Vector Space Model (VSM). The search results document is retrieved to the user based on the results of cosine similarity calculations between query vectors that have been expanded with vectors for each document. The document with the highest similarity value is placed at the top.

The value of the query vector and the document vector comes from the calculation of each weight. Each document will be calculated by the similarity of the query entered by the user using cosine similarity. It is shown in equation (1). For example based on Table III, the vector of query, document 1, and document 2 are (0,0,0.1598,0.1975,0), (0,0,0.1806,0.1806,0), (0,0,0.1505,0.1505,0). Based on the results of *cosine similarity* calculation in Table IV, document 1 has better degree of similarity compared to document 2. So, document 1 will be placed first.

Table III
Example of Query Weighting

Term	TF			idf	W=tf*idf		
	Q	d1	d2		Q	d1	d2
alam	0.531	0.6	0.666	0	0	0	0
bandung	0.5	0.6	0.666	0	0	0	0
ciburuy	0.531	-	-	-	-	-	-
cisanti	0.531	0.6	0.5	0.301	0.1598	0.1806	0.1505
floating	0.531	-	-	-	-	-	-
glamping	0.531	-	-	-	-	-	-
lakeside	0.531	-	-	-	-	-	-
lembang	0.562	-	-	-	-	-	-
market	0.531	-	-	-	-	-	-
patenggang	0.531	-	-	-	-	-	-
situ	0.656	0.6	0.5	0.301	0.1975	0.1806	0.1505
wisata	0.5	0.6	0.666	0	0	0	0

Table iv
Examples of Cosine Similarity Calculation Results between Queries and Documents

Cosine Similarity	d1	d2
	0,995	0,992

IV. Results And Discussion

The testing is done by counting and comparing the Mean Average Precision (MAP) value of fifty test queries entered by users between an ontology-based IR system and an IR system without ontology. When testing, the k document used for the calculation of MAP is the top ten documents retrieved by the system. From the ten documents displayed, every user is being asked to answer which document was relevant.The system test result is shown in Table V.

Based on the test results from fifty test queries in Fig. 6, the ontology-based search system have a higher MAP value of 0,563. While the search system that only applies the VSM method has a MAP value of 0,315. When the relevant document is not retrieved at all, the average precision value is 0 [3]. It means the system is not showing any relevant document at the top ten ranked documents. The position of the relevant document determines the result of the calculation, if the relevant document is in the top position without being cut off by the irrelevant document, the value is 1. From these results, we create more detailed analysis of queries with expansion and query without expansion.

We choose three queries based on class and individuals found on ontology, and query that are not found on ontology. The first query is '*Kuliner di bandung'*, the ontology stores information about individuals and types of culinary to be expanded, such as *cuanki serayu, nasi kalong, kuliner khas jawa barat*, etc. So, if the word 'kuliner' is not entered in the document, system will display the document containing other words from the results of the expansion. Unlike the VSM method without expansion from ontology, the document displayed is a document with '*kuliner'* only as the keyword.

The second query is *'Bukit Moko'*. Ontology stores information about *Bukit Moko* such as the address and rates. In addition, individuals who have the same class as *Bukit Moko* can be found. So that the results of the expansion are related to *Bukit Moko*. Whereas the non-expansion query only searches for documents based on the initial keyword.

Table V
Testing Result

Query	Average Precision		Query	Average Precision	
	Ontology+VSM	VSM		Ontology+VSM	VSM
Kuliner Bandung	0.96	0.331	tebing keraton	0.2	0.143
Penginapan di Lembang	0.75	0.278	pemandian air panas	0.892	0.917
kolam renang di bandung	0.416	0.756	tempat hits bandung	0.1	0.163
fasilitas glamping lakeside	0.531	0.881	farmhouse	0.734	0
harga tiket masuk Trans Studio Bandung	0.833	0.125	Wisata Danau	0.747	0.211
hotel bandung murah	0.679	0.111	restoran di bandung	0.722	0.417
tempat wisata di bandung	0.367	0,225	Kemah di Ranca Upas	0.81	0
tangkuban perahu	0.333	0.5	Puncak bintang	0.403	0.2
oleh-oleh bandung	0	0	daftar harga villa di bandung	0.767	0.278
curug layung	0.1	1	harga tiket kebun binatang bandung	1	0
wisata alam ranca upas	0.2	0.1	Maribaya Glamping	0.156	0.1
hiburan malam	0.225	0.2	Bukit Bintang	0	0
Wisata Keluarga di Lembang	0.833	0.667	Taman Hutan Raya	1	0.167
Dusun bambu	0.496	0.239	Villa di lembang	1	0.111
harga tiket dan jam operasional tahura	0.33	0.125	Situ patenggang	0.33	0,1
wisata alam	0.678	0.262	Lereng Anteng	1	0.243
wisata instagramable di bandung	0.251	0.517	Harga penginapan di hotel	0.25	1
kawah putih	0.458	0.583	Tempat kopi	0.73	0.242
Rekomendasi kuliner bandung	1	0	Situ Cileunca	1	0.111
art deco hotel	1	0.333	kebun teh	0.444	0.444
Hotel daerah dago	0.367	0.25	Kuliner Murah	0.81	0.267
wisata air	0.916	0.71	kolam renang indoor	0	0,5
jayagiri	0.778	1	10 lokasi terbaik wisata alam di bandung	0.815	0.616
curug cimahi	0.25	0	Bukit Moko	0.267	0
cafe enak buat belajar	0.2	0.333	harga tiket masuk kebun binatang bandung hari biasa	1	0

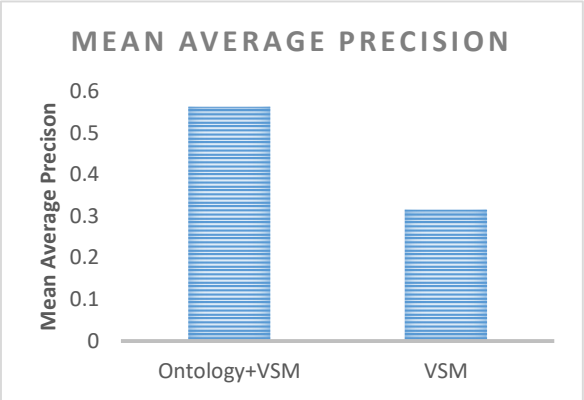


Fig. 6 Mean Average Precision of Test Result

The third query is *'tempat hits bandung'*. The ontology used has not information and knowledge about '*tempat hits di bandung'*, so there is no expansion concept. In addition, not all word in the queries are stored on an ontology-based index. This can affect the search results so that the MAP value in the third query is lower than the comparison method.

The ontology used is able to apply relevant information and knowledge to the concepts that cannot be applied using the VSM method. However, incomplete knowledge on ontology can affect search results. The VSM method is able to increase the number of documents given to users, both relevant and irrelevant documents. By using the VSM method, knowledge of concepts that have not been stored on an ontology, but contained in the user query will be given to the user if the concept is in the *index*.

V. Conclusion

In this research, ontology-based semantic search system with tourism around Bandung Raya as the domain is built to find out the influence of ontology in getting relevant information according to user needs. Ontology is used to expand users queries by extracting the information contained in the previous query concept. Besides ontology, the VSM method is used to represent the search. Based on the results of the testing, the system is able to provide relevant documents to users with an average MAP value better than the system that only uses the VSM method. The ontology-based search system have a higher MAP value of 0,563. While the search system that only applies the VSM method has a MAP value of 0,315.

For further research, ontology can be combined with other methods to get better results. In addition, knowledge or information on ontologies needs to be enriched such as adding other individual or class to meet the users' needs, that might add or change the information of tourism destination. This ontology can also be developed into a Bandung Raya tourism web so that it makes it easier for tourists to find tourist information when they want to visit Bandung Raya.

References

[1] W. Alromima, I. F., R. Elgohary, and M. Aref, "Ontology-based Query Expansion for Arabic Text Retrieval," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 8, pp. 223–230, 2016.

[2] T. Boonyoung and A. Mingkhwan, "Semantic ranking based on Computer Science Ontology weight," *2014 9th Int. Conf. Digit. Inf. Manag. ICDIM 2014*, no. February 2015, pp. 86–91, 2014.

[3] C. D. Manning, "Intro to Information Retrieval," *Inf. Retr. Boston.*, no. c, pp. 1–18, 2009.

[4] H. Dong, F. K. Hussain, and E. Chang, "A survey in traditional information retrieval models," *2008 2nd IEEE Int. Conf. Digit. Ecosyst. Technol. IEEE-DEST 2008*, no. March, pp. 397–402, 2008.

[5] D.Nandini, *Semantic Web and Ontology*. 2014.

[6] C. Paz-Trillo, R. Wassermann, and P. P. Braga, “An information retrieval application using ontologies,” *J. Brazilian Comput. Soc.*, vol. 11, no. 2, pp. 17–31, 2005.
[7] A. Roshdi and A. Roohparvar, “Review: Information Retrieval Techniques and Applications,” *Int. J. Comput. Networks Commun. Secur.*, vol. 3, no. 9, pp. 373–377, 2015.
[8] R. Suganyakala and R. R. Rajalaxmi, “Movie related information retrieval using ontology based semantic search,” *2013 Int. Conf. Inf. Commun. Embed. Syst. ICICES 2013*, pp. 421–424, 2013.
[9] J. M. Kassim and M. Rahmany, “Introduction to semantic search engine,” *Proc. 2009 Int. Conf. Electr. Eng. Informatics, ICEEI 2009*, vol. 2, no. August, pp. 380–386, 2009.
[10] B. Abu-Salih, “Applying Vector Space Model (VSM) Techniques in Information Retrieval for Arabic Language,” 2018.
[11] S. Shrivastava and S. Agrawal, “An Overview of Building Blocks of Semantic Web,” *Int. J. Comput. Appl.*, vol. 152, no. 10, pp. 17–20, 2016.
[12] L. Afuan and Azhari, “Penerapan Semantik Web Pada Ontologi Learning Resource Repositori,” *Semin. Nas. Teknol. Inf. dan Komun. 2016 (SENTIKA 2016)*, vol. 2016, no. Sentika, pp. 252–256, 2016.
[13] “Semantic Web.” [Online]. Available: https://www.w3.org/standards/semanticweb/.
[14] A. Nugroho, “Membangun Ontologi Jurnal Menggunakan Protege,” *J. Transform.*, vol. 10, no. 1, p. 20, 2018.
[15] Gita Desrianti, “Akurasi dalam Pencarian pada Search Engines,” *Makal. II2092 Probab. dan Stat.*, vol. 1, no. 1, p. 2011, 2011.
[16] M. Tang and Y. Sun, “Evaluation of Web-Based Search Engines Using User-Effort Measures,” 2003.
[17] D. Mardiana, “Tempat Wisata di Bandung.” [Online]. Available: https://tempatwisatadibandung.info/.