Semantic Search based on Ontology with Case Study: Indonesian Batik

Tri Kustanti Rahayu
Magister of Information System, Diponegoro University

Eko Sediyono
Magister of Information System, Satya Wacana Christian University

Oky Dwi Nurhayati
Magister of Information System, Diponegoro University

ABSTRACT

Many kind of searching technique has been developed today. One of these techniques is using ontology to support semantic technology. This technology is applied on this research. It is an experimental research to develop a semantic search based on ontology with case study on Indonesian article about batik. There are so many batik articles on the internet today. But the information about batik can not be easy to find. It because Indonesia is rich with batik. So many place produce batik with their own characteristic. The aim is to expand the meaning of keyword that user inputted. And the result shows that cosine value with expanded ontology is higher than without ontology. It almost double the value of cosine without ontology.

General Terms

Semantic, Ontology.

Keywords

TF/IDF, Vector Space Model, Document, Term, Ontology. Batik, Article

1. INTRODUCTION

Internet is used to publish many kind of information. It has been saving huge of information today. Conventional search based on string mostly failed to find relevant page and gives irrelevant feedback to show. Conventional search tend to find string that match with user input. Beside that how often a page is visited or accessed by user is also contributed to the search result. This shows that all data is available in the internet. User still find any difficulties to find it.

Many information techniques have been adopted to cover this problem included a technique that analyze the structure of the web. This technique counts the frequency of occurrence of keywords in the text description (such as title, body, anchor, text, and others). It can not be sure that the given result indicate a relation between the meaning and the keyword the user wanted [1]. Then came up the idea of semantic web by Tim Burners Lee to connect the web. Semantic search means searching the information on the web based on the meaning of the keyword. Which often not show in most of information search center. Beside that it requires complex programming.

Semantic is not only used to connect document from one to another but also to recognize the meaning of the information itself [2].

This research is focused on Indonesian article about batik. Batik is chosen because of knowledge of batik culture is faded away in Indonesia. Meanwhile the information of batik is spread so widely on the internet. People just do not know how to get it. The search engine today still can not show batik information precisely. It tends to show the commercial information about batik and ignore another informative side of batik.

2. RELATED WORK

Searching technique had been developed. Including full text search, metadata search, and semantic search. Semantic search attempts to improve the accuracy of search results by understanding the search intent and contextual meaning of terms as they appear in search data[3].

The idea of optimizing this semantic search by using the ontology framework is proposed in a research conducted by Soner Kara[4]. Another study was also conducted by Christian Lilik Henry. In his thesis, Christian mentions that the search accuracy resulting from ontology data is better than searching from simple data. The research is conducted on the travel and tourism ontology obtained precision value above 90% against 36.60% simple search data [5]. Other research based on ontology is also done by Agus Subhan Akbar[6]. This study used ontology to limit the domain of search results to tweets data to match the domains discussed. The result of 14,437 tweets, there are 13 tweets outside the domain of the discussion.

Another semantic search is done by Jihyun Lee. Its publication demonstrates effective ranking techniques as well as search techniques that emphasize on ontology relationships. The weighting measurement for semantic relation is ranked by considering a more meaningful relation between resource and keyword. To improve efficiency, shortening or trimming the search space using the length and weight threshold of the semantic relation [7].

In addition, research on batik ontology was previously conducted in 2010, by Syerina Azlin Md Nasir who examined the approach to integrate ontology through Knwoledge Management System (KMS). Syerina Azlin Md Nasir developed a knowledge model for mapping process between local ontologies and other ontologies to build ontology of batik culture [8].

3. RESEARCH METHODOLOGY

This experimental research is done by building a semantic application based on ontology using indoensian article about batik. The following methods and research stages are as follows:

3.1 Tf/Idf

The similarity of two words can be measured by assigning weight values to each word. Word weight is calculated based on the frequency of word occurrence in a document. The weight of term t in a document is obtained by multiplying the value of the term frequency by inverse document frequency.

$$W_{ij} = tf_{ij} \times idf_j$$

$$W_{ij} = tf_{ij} \times log\left(\frac{D}{df_j}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Caption::
W_{ij} : the weights of term t_j against the document d_i
tf_{ij} : number of occurrences of term t_j in a document d_i
D : the number of all documents in the database
df_j : Number of documents containing term t_j
(At least one word is term t_j)

3.2 Vector Space Model

The distance between the documents is measured through the resulting angle (Cosine). The angular similarity between the document vector and query shows similarities.

$$R(Q,D) = \cos\theta = \frac{Q.D}{|Q||D|} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ (2)$$

Caption:
Q : query weight
D : document weight
$|Q|$: distance query
$|D|$: distance document

3.3 Ontology

Ontology is built to support semantics. Ontology makes an unstructured domain to be structured. The ontology structure of batik that is built is as the figure bellow. Batik is an ancient product of people in Indonesia. The way of producing batik makes batik is valuable. There are many areas in Indonesia producing batik. It makes differentiation call between one and other. Beside that the design motif is influence also. It makes Indonesia rich with batik. But not many people understand about these richness.

The structure of batik domain can be described as follow:

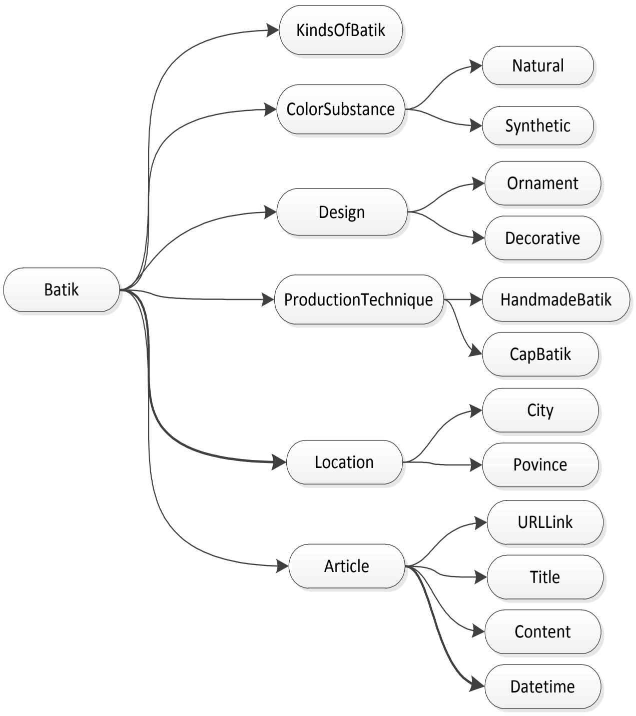


Figure 1. Batik structure ontology

3.4 Research Procedure

This research aims to build semantic web application based on ontology through the following stages:

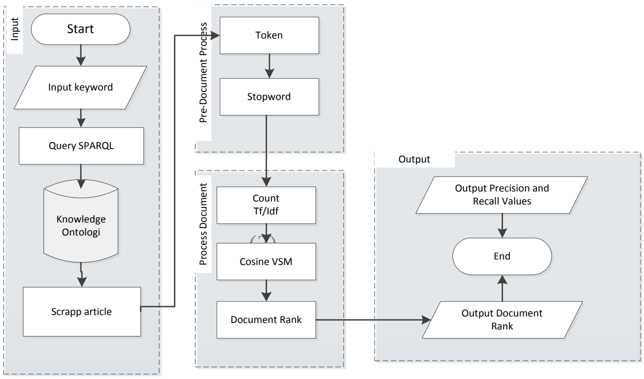


Figure 2: Flowchart system

This research will be divided into several stages as in the figure 2.

1. The first stage starts from inputting keywords by the user. Then the keyword will be expanded through sparql query to ontology. After that the system will scrap the article.
2. Phase pre-processing documents. At this phase, the articles that have been collected will then go through the process of tokenization and stopword to get the word base.
3. The third stage is counting process. At this stage, some calculations are performed to obtain cosine similarity values such as word frequency calculation, word weighting and count cosine value between query and documents. And after that the article can be ranked, the value of cosine that tends to show the document match with the query.

4. RESULT

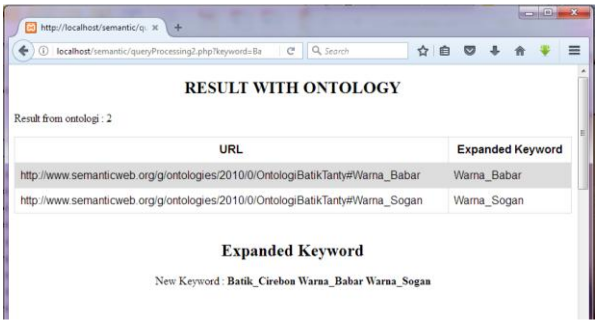
Conducted several times trial on data in the form of batik articles. Some results of the stages of research procedure as follows:

1. The first result is expanding user keyword meaning through sparql query. The figure 3 shows that user input keyword: *Batik Cirebon* is expanded and become more keywords such as *Batik Cirebon*, *Warna Babar*, *Warna Sogan.* This is a result from sparql query to ontology. The idea is to read the keyword from user input, after that extract the input so ontology can understand the input. And then ontology will process the keyword by the loqical query as follow.

```
$querystring = 'PREFIX batik:
<http://www.semanticweb.org/g/ontologies/201
0/0/OntologiBatikTanty#>
     SELECT ?Keyword2
      WHERE { batik:'.$sliceKeyword.' ?p
         ?Keyword2
      FILTER REGEX
         (str(?p),"OntologiBatikTanty")}';
```

This query shows `?Keyword2` as the expanded keyword, and shows all instance where the condition is batik ontology has an instance like user input `$sliceKeyword`. And filter function is for filtering string result on the predicate.

The result is shown as follow.



http://localhost/semantic/q
localhost/semantic/queryProcessing2.php?keyword=Ba
Search

RESULT WITH ONTOLOGY

Result from ontologi : 2

URL	Expanded Keyword
http://www.semanticweb.org/g/ontologies/2010/0/OntologiBatikTanty#Warna_Babar	Warna_Babar
http://www.semanticweb.org/g/ontologies/2010/0/OntologiBatikTanty#Warna_Sogan	Warna_Sogan

Expanded Keyword

New Keyword : Batik_Cirebon Warna_Babar Warna_Sogan

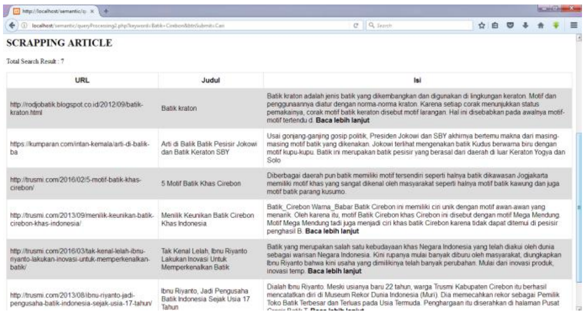
Figure 3. Result with ontology

After get the expanded keyword from ontology, the system continue to scrap article from mysql using the code below.

```
$keywordQuery = "SELECT url, judul, isi FROM
artikel ";
     if (sizeof($splitKeyword) > 0){
     for ($i = 0; $i < sizeof
         ($splitKeyword); $i++){
      if ($i == 0) {$keywordQuery =
     $keywordQuery . " WHERE isi LIKE
     '%$splitKeyword[$i]%' ";
              }else {
$keywordQuery = $keywordQuery . " OR isi
     LIKE '%$splitKeyword[$i]%'";
}
}
```

The query is to select `url`, `judul` and `isi` from `article` table where the condition is `isi` has string like in variable `$splitKeyword[$i]`.

And the result is shown in figure 4.



http://localhost/semantic/q
localhost/semantic/queryProcessing2.php?keyword=Batik+Cirebon&btnSubmit=Cari
Search

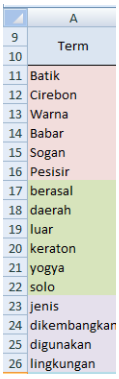
SCRAPPING ARTICLE

Total Search Result : 7

URL	Judul	Isi
http://rodjobatik.blogspot.co.id/2012/09/batik-kraton.html	Batik kraton	Batik kraton adalah jenis batik yang dikembangkan dan digunakan di lingkungan keraton. Motif dan penggunaannya diatur dengan norma-norma kraton. Karena setiap corak menunjukkan status pemakainya, corak motif batik keraton disebut motif larangan. Hal ini disebabkan pada awalnya motif-motif tertendu d. **Baca lebih lanjut**
https://kumparan.com/intan-kemala/arti-di-balik-ba	Arti di Balik Batik Pesisir Jokowi dan Batik Keraton SBY	Usai gonjang-ganjing gosip politik, Presiden Jokowi dan SBY akhirnya bertemu makna dari masing-masing motif batik yang dikenakan. Jokowi terlihat mengenakan batik Kudus berwarna biru dengan motif kupu-kupu. Batik ini merupakan batik pesisir yang berasal dari daerah di luar Keraton Yogya dan Solo
http://trusmi.com/2016/02/5-motif-batik-khas-cirebon/	5 Motif Batik Khas Cirebon	Diberbagai daerah pun batik memiliki motif tersendiri seperti halnya batik dikawasan Jogjakarta memiliki motif khas yang sangat dikenal oleh masyarakat seperti halnya motif batik kawung dan juga motif batik parang kusumo.
http://trusmi.com/2013/09/menilik-keunikan-batik-cirebon-khas-indonesia/	Menilik Keunikan Batik Cirebon Khas Indonesia	Batik_Cirebon Warna_Babar Batik Cirebon ini memiliki ciri unik dengan motif awan-awan yang menarik. Oleh karena itu, motif Batik Cirebon khas Cirebon ini disebut dengan motif Mega Mendung. Motif Mega Mendung tadi juga menjadi ciri khas batik Cirebon karena tidak dapat ditemui di pesisir penghasil B. **Baca lebih lanjut**
http://trusmi.com/2016/03/tak-kenal-lelah-ibnu-riyanto-lakukan-inovasi-untuk-memperkenalkan-batik/	Tak Kenal Lelah, Ibnu Riyanto Lakukan Inovasi Untuk Memperkenalkan Batik	Batik yang merupakan salah satu kebudayaan khas Negara Indonesia yang telah diakui oleh dunia sebagai warisan Negara Indonesia. Kini rupanya mulai banyak diburu oleh masyarakat, diungkapkan Ibnu Riyanto bahwa kini usaha yang dimilikinya telah banyak perubahan. Mulai dari inovasi produk, inovasi temp. **Baca lebih lanjut**
http://trusmi.com/2013/08/ibnu-riyanto-jadi-pengusaha-batik-indonesia-sejak-usia-17-tahun/	Ibnu Riyanto, Jadi Pengusaha Batik Indonesia Sejak Usia 17 Tahun	Dialah Ibnu Riyanto. Meski usianya baru 22 tahun, warga Trusmi Kabupaten Cirebon itu berhasil mencatatkan diri di Museum Rekor Dunia Indonesia (Muri). Dia memecahkan rekor sebagai Pemilik Toko Batik Terbesar dan Terluas pada Usia Termuda. Penghargaan itu diserahkan di halaman Pusat

Figure 4. *Scrapping* article

2. Pre-Document Process Result
   On this stage there are two processes to get the term base. The first is to divided the article into strings. And then to remove unnecessary word as in the stopword list.



	A
9	Term
10	
11	Batik
12	Cirebon
13	Warna
14	Babar
15	Sogan
16	Pesisir
17	berasal
18	daerah
19	luar
20	keraton
21	yogya
22	solo
23	jenis
24	dikembangkan
25	digunakan
26	lingkungan

Figure 5. Term base

3. Document Process Result
   Documents that have been through pre document process will be counted the weight and the cosine value. The weight is counted using first formula above. This process considers that every term is important. That is why every term is counted its frequency (f). Then, to anticipate huge document frequency it needs to be normalized before. To normalize the frequency is to divide its frequency with total frequency from all document. So that the value will be in between 0 to 1. As shown in table 1 that the term frequency of the query and term frequency in all documents is counted. Then, these value used to count inverse document frequency (idf).

Table 1. Term Frequency and Inverse Document Frequency

	Frequency (f)						Document frequency (df)	Inverse df
Terms	**Q**	**D1**	**D2**	**D3**	**D4**	**D5**		**=Log(n/df)**
Batik	0.166667	0.2		0.142857	0.166667		0.509524	0.769956772
Cirebon	0.166667			0.142857	0.166667		0.309524	0.986427193
Warna	0.333333	0.2		0.142857	0.166667	0.2	0.67619	0.647052205
Solo		0.2				0.2	0.2	1.176091259

After that is to count the weight using Tf/Idf method with formula (1). Weighting value is to multiply between the term frequency and the inverse document frequency. And the result as shown in table 2 below.

Table 2. Term Weight

Weight (tf*idf)					
Q	**D1**	**D2**	**D3**	**D4**	**D5**
0.128326	0.153991	0	0.109994	0.128326	0
0.164405	0	0	0.140918	0.164405	0
0.215684	0.12941	0	0.092436	0.107842	0.12941
0	0.235218	0	0	0	0.235218

Next is to count the cosine value using the second formula above. And the result is shown on Table 3. Cosine value shows the angular similarity between query (Q) and documents (D). The bigger the value shows the more similar the query with the documents. And the documents with 0 value such as in the second experiment means that the query is not similar with the query which is there is no keyword found in document 2 (D2). And the most similar document is the fourth experiment, query with document 4 (D4) which the cosine value is 0.646316.

Table 3. Cosine Value

Cosine Similarity	
(Q,D1)	0.313924
(Q,D2)	0
(Q,D3)	0.550374
(Q,D4)	0.646316
(Q,D5)	0.300922

5. DISCUSSIONS

The use of ontology to support semantic technology is used widely today. Some researcher use ontology to raise the accuracy value like Christian done [5]. And other is to limit the domain discussed so the search result value is higher [6]. Meanwhile this experiment using ontology to expand the search query. The idea is to take the advantage of ontology characteristic that can make an unstructured domain to be structured. So that the information about a domain can be clearly visible. As happened to Indonesian batik which has so many kinds and names. Ontology helps to make it structured. Then it can be expanded. So that the information could be more specific. For example input query like *Batik Keraton*, then system will show additional information that *Batik Keraton* comes from the palace area such as *Solo*, *Yogya, Cirebon* and *Sumenep*. Also that Batik Keraton has special design like design *Parang Barong*, *Parang Rusak*, *Udan Liris* and so on.

From the experiment shows on the table below, the cosine value with ontology is higher than cosine value without ontology. it because expanded keyword gives more value to a similar document. The average value is twice from the result cosine value without ontology. Document with no similarity with the query at all shows the cosine value 0. Beside that it is shown on the experiment that document still has cosine value (similarity) with the keyword eventhough keyword is not show up in the document. It is show on the fifth document on the second experiment.

Table 4. Result Experiment

Experiment	Article	Cosine Value	
		With Ontology	Without Ontology
1	D1	0.261642512	0.151735083
	D2	0	0
	D3	0.489950783	0.295073669
2	D1	0.313923634	0.12777827
	D2	0	0
	D3	0.550373574	0.295073669
	D4	0.646316053	0.443016428
	D5	0.300922238	0

6. CONCLUSIONS

This experimental research is studying about semantic technology. By using the characteristic of ontology can support semantic web technology. Ontology makes unstructured domain became structured. And many developers has take the advantages from this. One of this way is to expand keyword that inputted by user on a search engine. And it shows that the cosine value result with expanding keyword from ontology is higher than keyword without ontology. And document with no keyword in it could have similarity value because of this expanding technique also. Another technique also can be developed to support semantic technology such as improving knowledge of ontology such as adding picture as entity of ontology. So that the information scope can be wider. And also that ontology can be more powerful as the basic of searching technique to face big data era.

7. REFERENCES

[1] Li Y., Wang Y., Huang X, 2007, *A Relation Based Search Engine In Semantic Web*, IEEE Transaction on knowledge and data engineering 19(2), 273-282 (February)

[2] Hebeler John, Matthew Fisher, Ryan Blace, Andrew Perez Lopez, 2009, *Semantic Web Programming*, Wiley Peublishing Inc, Indianapolis, Indiana

[3] Sarno Riyanarto dan Rahutomo Faisal, 2008, "*Penerapan Algoritma Weighted Tree Similarity Untuk Pencarian Semantik Wikipedia*", JUTI volume 7, Nomor 1, Januari 2008: 35-42

[4] Kara Soner, Ozgur Alan, Orkunt Sabuncu, Samet Akpinar, Nihan K. Cicekli, Ferda N Alpaslan, 2012, *An Ontology Based Retrieval System Using Semantic Indexing*, International Journal of Information Systems 37 (2012) 294-305

[5] Christian, Lilik Henri, 2014, *Pencarian Informasi Berbasis Ontologi Menggunakan Semantik Indexing Pada Website Jejaring Sosial*, Masters Thesis, Diponegoro University, Semarang

[6] Akbar Agus Subhan, 2015, *Analisis Sentimen Berbasis Ontologi di Level Kalimat Untuk Mengukur Persepsi Produk*, Masters Thesis, Diponegoro University, Semarang

[7] Lee Jihyun, Jun-Ki Min, Alice Oh, Chin-Wan Chung, 2013, "*Effective Rangking and Search Techniques for Web Resources considering Semantic Relationships*", International Journal of Information Processing and Management 50 (2014) 132-135

[8] Syerina Azlin Md Nasir, Nor Laila Md Noor, 2010, *Integrating ontology based approach in Knowledge Management System (KMS): Construction of Batik Heritage Ontology*, International conference of science and social research.