Journal of Engineering and Applied Sciences 12 (8): 1979-1982, 2017
ISSN: 1816-949X

Semantic Similarity for Search Engine Enhancement

[1]Detty Purnamasari, [2]Lintang Yuniar Banowosari, [3]Reni Diah Kusumawati,
[2]Dessy Wulandari Ap and [4]Disa Restu Kusuma
[1]Department of Information System, Gunadarma University,
Jl. Margonda Raya No. 100 Pondok Cina, Depok, Indonesia
[2]Department of Information Management,
[3]Department of Accounting, [4]Technical Information,
Gunadarma University, Jakarta, Indonesia

Abstract: Now a days we still found the search results of search engines on websites that are not in accordance with the wishes of user and only provide the information that the keyword search could not be found. For e-Commerce website, this can cause the website would be left by users/prospective buyers so based on that, this research is to increase search results on search engine from e-Commerce websites using semantic similarity and query rewriting. Semantic similarity used is engineering calculations by Leacock and Chodrow. The illustrations in this research is to provide a snapshot query rewriting to make a new query results from semantic similarity and the test method is carried out on a prototype e-Commerce website as well as provides results that the search engine on the website after added semantic similarity approach and query rewriting that provide a better search results.

Key words: e-Commerce, search engine, semantic similarity, query rewriting, website

INTRODUCTION

Websites can be thought as a book with many pages accessed via internet. Nowadays, utilization of website is in all fields such as make a sale via online, known as e-Commerce. Search engines which provided by e-Commerce website is aim to make easier for users/visitors of website in search of products/items to be purchased but the current condition, the search engine still give the results not in accordance with the wishes of user or even informed that the product/goods sought was not found in e-Commerce website. The storefront of a virtual store is the website, if the website gets high marks in usability, the website accepted and used productively is very high (Imran and Sharan, 2009). Therefore, usability storefront becomes a major determining factor in customer acceptance of virtual stores.

This study use semantic similarity approach and query rewriting embedded in the search engine of e-Commerce website to improve search results that match with the user desires/prospective buyers. Semantic similarity used is the calculation by means Leacock and Chodorow which calculates the value of semantic similarity based on a path length of two words.

Literature review: According to Turban (2004), e-Commerce is a process of buying, selling, transfer or exchange of products, services and information via computer networks including e-Internet. e-Commerce in Indonesia has existed since 1996 with the establishment of Dyviacom Intrabumi or d-net as the pioneer of online transactions. On the website, points to consider are:

- Features website, completeness features of the website will make consumers interest to get back on online shopping sites because consumers feel comfortable in doing a product search (Pentina *et al*., 2011)
- Content website, a variety of stimuli that can affect consumers such as the factor of product presentation and interactivity on the website (McCormick and Livett, 2012)
- Leisure website, online shoppers try and adopt the Internet to make purchases for their comfort at the retailer's website (Lee *et al*., 2011)

MATERIALS AND METHODS

Semantic similarity: Semantic similarity is a problem on the semantic relationship. Semantic relationships is an

Corresponding Author: Detty Purnamasari, Department of Information System, Gunadarma University,
Jl. Margonda Raya No. 100 Pondok Cina, Depok, Indonesia

approach to find out how the relationship between these two concepts in the use and relationships. The relation between the concepts is not always symmetrical, if the 2 concepts are the same, it means also has relationships, but if there is a relationship it doesn't mean the same (Wicaksana and Wayan, 2004). Semantic similarity of the two words/concepts is represented in the form of value. The calculation of the value can be done with this approach). Path length which is the technique Leacock-Chodorow, Resnik, Wu-Palmer). Information content using Lin and Jiang Conrath technique. In this study, the value of semantic similarity obtained by Leacock-Chodorow techniques, the following equation is:

$$lch = \log\left(\frac{2\times D}{(length(c1, c2))}\right) ...F.1$$

Where:

c1 = Concept 1
c2 = Concept 2
(c1, c2) = The shortest path length (i.e., minimum number edge between 2 concepts)
D = The maximum depth of taxonomy (the maximum number of nodes of scheme ontology of the 2 concepts)

Here are example for calculation of semantic similarity value using techniques Leacock-Chodorow. Word1 (c1) is "teacher" and word2 (c2) is "employee":

- Find length for c1 (teacher) and c2 (employee)
- Input c1 and c2 in wordNet
- Calculate to the length from entity until c1 or c2
- We get value for c1 is '9' and c2 is '7'. Length value for c1 and c2 is the minimum value and for this example, length value is '7'
- We get D value by comparing the number of nodes of c1 and c2, the maximum number of nodes delivered c1 and c2 is '10'
- Value of semantic similarity for word1 "teacher" and word2 "employee" is 0.45

Query rewriting: According to Shiri and Revie (2006), query rewriting is the stage of the information search process in which the initial query statement users enhanced by adding a search term to improve the performance of information retrieval. Query rewriting is also termed as query expansion which according to Imran and Sharan (2009) is the process of completing additional term or phrase in the beginning of query as a way to improve the performance of information retrieval. The key point of query rewriting is how to get the best improvements words used to expand the initial query (Imran and Sharan, 2009).

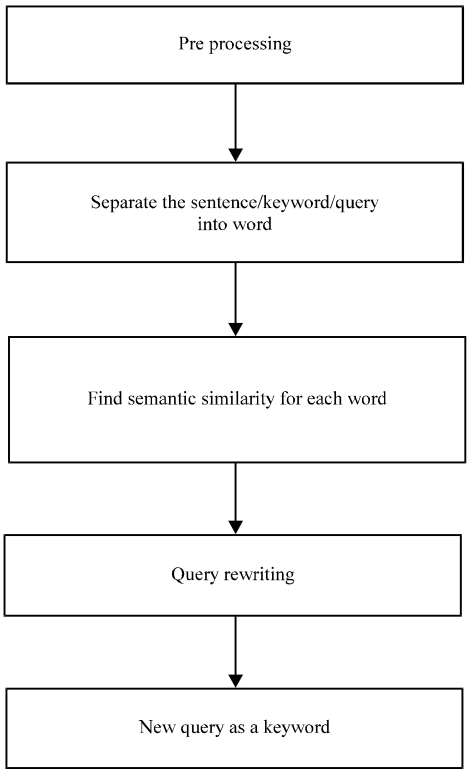

Fig. 1: Steps for research

Semantic similarity and query rewriting method: Semantic similarity is one of way that can be used to improve search results based on keywords/query desired by website users. Stages in Fig. 1 is preceded by preprocessing which is build a database of 'dictionary' and the database 'similarity'. Database 'dictionary' contains words in Indonesian and English which are used as tools in the search for semantic similarity value of word net. Database 'similarity' contains words and other words that have the same meaning with the values obtained from word net with the calculation of leacock-chodorow. The next step is:

- If a keyword is a phrase/more than one word, then do separation word of the sentence
- The results of the separation sentence by the word of Step 2, then find the similar meaning by using a database of 'similarity' based on the values that have been entered by user
- The word with has the similar meaning will be reunited into a new sentence/new query

The process of separating the sentence is as follows:

- Read the query entered by user
- Read the query by one character

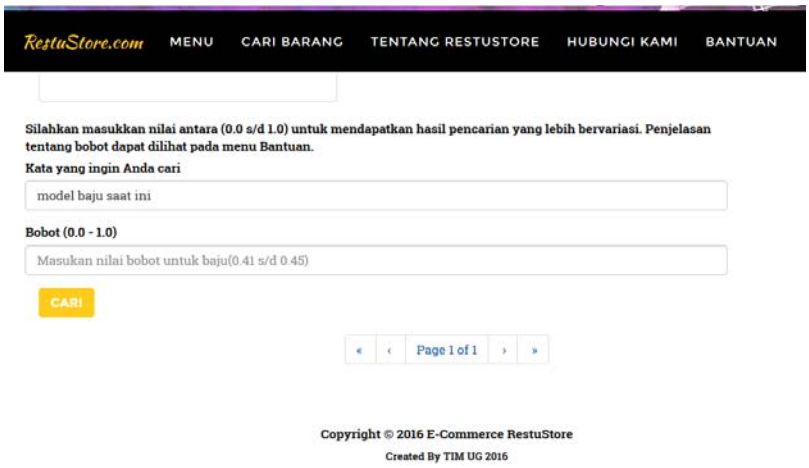

Fig. 2: e-Commerce website for testing

Table 1: Sample data in database 'similarity'

Id	Word 1	Word 2	Values
1	Clothes	Dress	0.41
2	Clothes	Shirt	0.41
3	Clothes	T-shirt	0.45
4	Children	Kid	0.35
5	Pants	Jeans	0.39

- If the character found is space then mark it as the n word that has been discovered and save
- Return to Step 3 and ended up characters which is read a period (.) or spaces 5 times

Having obtained a new word which is the search result of the equivalence meaning, then the word will be in the chain back into a new query (query rewriting), illustrated as follows: Key word search/query = k(1)+ k(2)+k(3) suppose the results of equivalence meaning/semantic similarity by using a database of 'similarit' is obtained:

- k(1) = k(1a); k(1b)
- k(2) = k(2a)

then query rewriting is done will result in a new query as follows:

- k(1)+k(2a)+k(3)
- k(1a)+k(2)+k(3)
- k(1a)+k(2a)+k(3)
- k(1b)+k(2)+k(3)
- k(1b)+k(2a)+k(3)

RESULTS AND DISCUSSION

Prototype of e-Commerce websites include search engine facikity which used in this study is to conduct testing. Figure 2 shows one page of an e-Commerce website which shows the page to search the product (Stahl, 1987). On the page of e-Commerce websites with search engine, user enters a weight between 0.1 up to 1.0 as the value to look for the equivalence meaning to the keywords entered by the user.

Table 1 is an example of data contained in a database of 'similarity' which used for the process of finding new words that will form the keyword/new query. Research conducted in this study is designed to recognize key words in Indonesian which used when do searching so the illustration is given by using Indonesian language:

- Example
- Keywords entered
- Model baju saat ini
- Clothes model nowadays

These keywords, using break the sentence algorithms into words, got 4 words such as:

- Model (model)
- Baju (clothes)
- Saat (today)
- Ini (this)

Having obtained per word, using a database of 'similarity', find the other words that have similar meaning with keywords. Suppose a user using a similarity score is 0.45, then for the word 'baju (clothes)' obtained the other words that have similar meanings are 'kaos (T-Shirt)'. Then, the word obtained by searching the similar meaning will be reunited into a new query and the newly formed query examples are: "model kaos saat ini"/T-shirt model for today. Test of search engines which is given semantic similarity approach and query rewriting on prototype e-Commerce website built gives the results shown in Table 2.

Table 2: Sample test results

Query	Values	New query	Search result
Model baju saat ini	0.2	Model kaos saat ini	9 item product
Clothes model now a days		Shirts model now a days	
		Model kemeja saat ini	
		Model shirt today	
		Model gaun saat ini	
		Model dress this time	
Model celana saat ini	0.2	Model jins saat ini	7 item products
Pants model now a days		Model jeans today	
		Model celana panjang saat ini	
		Model trousers today	
		Penddek saat ini	
		Model shorts today	

Query used as the keyword is the first queries used in the search which did not give the search results and in Table 2 are shown that after using the method of semantic similarity and query rewriting has formed a new query and search results no longer null but give the results search as much as the amount listed on the 'search.

CONCLUSION

Semantic similarity and query rewriting approach which is added to search engine may improve search results, because by using this method as an indicator of a search query into more than one query and the meaning contained in the new query would be the same. Semantic similarity approach in this study can be used for keywords other than Indonesian by enriching the database of 'similarity' with other languages and to learn another language sentence structure to query rewriting.

Advanced research related to this topic is to search for methods to build a database of 'similarity' that can be automated reference to word net and enriched with semantic similarity value using calculation techniques other than Lecock and Codorow such as the calculations base on information content.

ACKNOWLEDGEMENTS

This research is partially supported by Gunadarma University, Research Institution at Gunadarma University Jakarta Indonesia, Hibah Penelitian Unggulan Perguruan Tinggi 2016.

REFERENCES

Imran, H. and A. Sharan, 2009. Thesaurus and query expansion. Intl. J. Comput. Sci. Inf. Technol., 1: 89-97.

Lee, C.H., C.U. Eze and O.N. Ndubisi, 2011. Analyzing key determinants of online repurchase intentions. Asia Pac. J. Marketing Logist., 23: 200-221.

McCormick, H. and C. Livett, 2012. Analysing the influence of the presentation of fashion garments on young consumers' online behaviour. J. Fashion Marketing Manage. Intl. J., 16: 21-41.

Pentina, I., A. Amialchuk and D.G. Taylor, 2011. Exploring effects of online shopping experiences on browser satisfaction and E-tail performance. Intl. J. Retail Distrib. Manage., 39: 742-758.

Shiri, A. and C. Revie, 2006. Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation. J. Am. Soc. Inf. Sci. Technol., 57: 462-478.

Stahl, B., 1987. Testing for usability can head off disaster. Computerworld, 21: 83-92.

Turban, E., 2004. E-Commerce, a Managerial Perspective. 4th Edn., Prentice Hall, Inc., New Jersey, USA.,.

Wicaksana, I. and S. Wayan, 2004. Survey and Evaluation of Ontology Development Method. Gunadarma University, Depok, Indonesia,.