Concept Similarity Searching for Support Query Rewriting

Detty Purnamasari, Lily Wulandari, Ahmad Muhammad Thantawi, and I Wayan Simri Wicaksana

***Abstract*—Internet search engines are used for the search process by entering the keywords/query. Search performed by the user information by using keywords/query is usually a string of words/concepts (sentences), and search engines provide results do not necessarily match those in intent by users. Keywords/query that used by the user to the search process can be developed into a new query using query rewriting. Query rewriting is a way to generate new queries and has a similar meaning to the keyword used in the initial query, and the results are given in the search engines to find information closer to the user's taste. Query rewriting can be done using the following steps: i). extraction query into word/concept, ii). search for word concepts similarity, iii). Preparation of a new query from the concept similarity (query rewriting). In this article, the approach is developed to search the concept similarity in the second phase of the process of query rewriting is performed after extraction queries. Phase of extraction query is to break the query into a word/concept. Queries used in this approach is the query in Indonesian. The illustrations in this article are given to provide a clear picture of the approach in the search for equivalence concepts to support query rewriting.**

***Index Terms*—Concept similarity, information searching in internet, search engine, query rewriting.**

I. Introduction

Information can be easily obtained via the Internet. Information search on the Internet can be done by utilizing the application of existing search engines, like Google, Yahoo, and Bing. The search engine can be regarded as a device that is used to search for information in a collection of documents, and the user simply enter a keyword of the information sought and within a relatively short period of time the system will display a list of documents according to the user's information needs [1].

Google is one of search engine information on the Internet that popular today, the reason is because Google has a simple display and easy to use, but it can also provide Google search results with many URLs [2].

Information search on the Internet using keywords/query does not always give accurate results or does not match with keywords that user used to searching. This is because search engines are not able to look for patterns of relevant documents or the user does not state when the search request with the correct information [1]. The lack of a search engine can be overcome by doing a query rewriting.

Query rewriting [3] are the stages of the process of information retrieval by user's initial query statement by adding enhanced search terms to improve information retrieval performance. Query rewriting can also be referred to as query expansion, which according to Hazra Imran *et al*. [4] is the process of completing additional query term or phrase in the beginning as a way to improve the performance of information retrieval. The key point of query rewriting is how to get the best repair words that are used to expand the initial query [4].

Zhu Kunpeng, *et al.* [5] do query expansion begins with measuring the quality of the query by using the ambiguity analysis, and increased the term by using query log mining.

Fig. 1 is a query rewriting stage that can be done to enrich keywords/query that is used in the search for information on the Internet. In the research conducted, the query is a query that is used in Indonesian, and the preprocessing performed before entry into the query rewriting stage. Preprocessing is done by making two (2) databases, namely: i). “dictionary” database containing words in Indonesian and English, ii). “similarity" database that contains words with his equivalence and similarity weights were obtained from WordNet.

Query rewriting process is the process of making other queries based on concepts used in the initial query. Thus it is necessary other concepts with the same meaning to the concept of the initial query to build other queries. Contribution of this research is concept similarity searching that have the same meaning. This process can enrich information retrieval. This process is an important process because the results of this process will be used to form other queries that have the same meaning as the initial query.

For example, if the initial query searched for "number of cars in Jakarta", if the search engines on the internet without enriched with equivalence word, then it just searched for "number of cars in Jakarta". If the query process is preceded by a query rewriting process enriched by concept similarity searching then the information retrieval instead of just using "number of cars in Jakarta" but also another query statement such as "A lot of cars in Jakarta", and "total car in Jakarta". An example of this is the result of the query rewriting if only the concept of "number" which sought equivalence. The process of query rewriting can produce query more than that if the

Manuscript received January 9, 2015; revise May 20, 2015. This work was supported in part by Gunadarma University Jakarta and YAI Jakarta Indonesia.

Detty Purnamasari and Lily Wulandari are with the Department of Information System, Gunadarma University. Jl. Margonda Raya No. 100 Pondok Cina Depok, Indonesia (e-mail: detty@staff.gunadarma.ac.id, lily@staff.gunadarma.ac.id).

Ahmad Muhammad Thantawi is with the Department of Informatics Engineering, University of Persada Indonesia YAI-Jakarta, Indonesia (e-mail: thantawi@yahoo.com)

I Wayan Simri Wicaksana is with the Department of Information Technology, Gunadarma University. Jl. Margonda Raya No. 100 Pondok Cina Depok, Indonesia (e-mail: iwayan@staff.gunadarma.ac.id).

other concepts forming the initial query also searched equivalence concept.

Stages of a query rewriting are:

1) Perform extraction query into word/concept
2) Change the concept into English using the "dictionary" database
3) Finding concept similarity by using a "similarity" database. The user can be specify the desired weight similarity value.
4) Change back the concept and the results of phase 3 into Indonesian by using a database of "dictionary"
5) Perform query formulation based on the concept of results stage 4.

Semantic Similarity is an issue on the semantic relationships. Semantic relationship is an approach to determine how the relationship of two concepts in use and their relationships, despite some similarities only consider the IS-A relationship (hyponymy). Relationships between concepts are not always symmetrical, if the two concepts are the same, meaning also has a relationship, but if the same does not necessarily mean related [6].

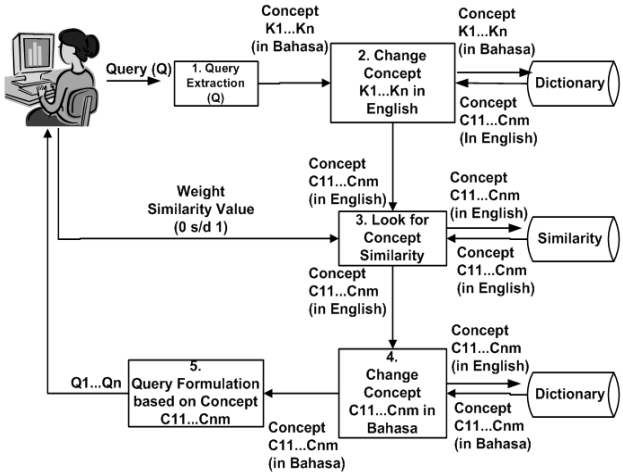

Fig. 1. Query rewriting stages.

Semantic Similarity calculation is a process that requires the involvement of several disciplines, such as language, computer, mathematical logic and the domain in question. The initial step is the calculation of semantic similarity refers to the similarity of terminology is often called a label, and is referred to as matching labels. Terminology that may encompass a class (classes), properties (property) to cases (instances) [7].

In this article is developed approach to search the concept similarity that is part of the overall process steps to perform query rewriting (as shown in Fig. 1. Similarity of concept search is on step 3). Similarity of concept search performed after query extraction that is the first step of the overall process of query rewriting.

II. Preprocessing to Support Concept Similarity Search

The approach is developed to search concept similarity is supported by preprocessing. Preprocessing performed by the establishment of a "dictionary" database, and "similarity" database.

A. "Dictionary" Database

On the establishment of a "dictionary" database, the steps being taken are:

1) Conducted a survey to find the word that is often used in the search for information on the Internet.
2) Translating words into English produced by an online dictionary which is then stored in the database Indonesian-English dictionary.
3) Finding equivalence word in English translation obtained by using a thesaurus dictionary online.
4) Translate the words results from 3 phase to Indonesian process using an online dictionary which is then stored in the database Indonesian-English dictionary.

B. "Similarity" Database

Establishment of a "similarity" database using tools to measure the similarity of the two words in the English language, namely WordNet similarity (http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi), which was developed by Ted Pedersen and Jason Michelizzi. In similarity calculation can be done by eleven ways, some of which are Path Length, Leacock & Chodorow, Wu & Palmer, Resnik, and Jiang & Conrath.

Calculation of similarity with the Jiang & Conrath (JNC) showed the best results [8], [9], so in this study used the method of Jiang and Conrath (JNC) to find the weights measure the similarity as a value for the equivalence.

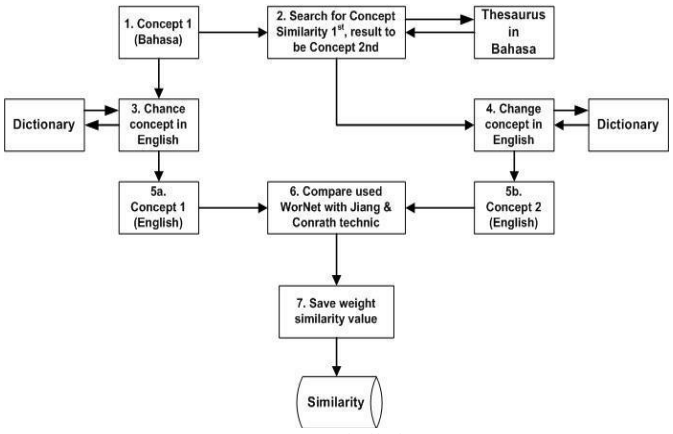

Fig. 2. Steps to build "similarity" database.

Fig. 2 is the steps being taken to create "similarity" database. The steps are:

1) Take one concept in Indonesian language.
2) Equivalence of a concept searched using the Indonesian thesaurus
3) Then the original concept in Indonesian converted into English by using the dictionary.
4) The concept that is found from the results of the search process concept similarity also changed into English.
5) (a) and (b) is the result of the conversion of the initial concept and the concept similarity has been the English language.
6) Both concepts that has changed is inserted into WordNet and compared using the technique of Jiang and Conrath.
7) The output of the WordNet is similarity weight values. Concept, the equivalence of concept and the resulting

weight of WordNet is stored into the "similarity" database.

III. Search Approach for Concept Similarity

Steps being taken to search similarity of concept is done by comparison of similarity weight gained during the search for equivalence of concepts from database (BSimD) with similarity weight is entered by the user (BSimP).

Fig. 3 is the stage for the search concept similarity. The concept used is the result of the query extraction (*C*11 ... CNM) that have been changed to English. The equivalence of concept will be searched in the "similarity" database. If the concept is found, it will be known concept similarity and also known similarity weight (obtained at the stage of pre-processing by utilizing WordNet). Concept of the initial query, concept similarity, and the weight of similarity (BSimD) will be saved. But if it is not found, then it will conduct a search to the next concept.

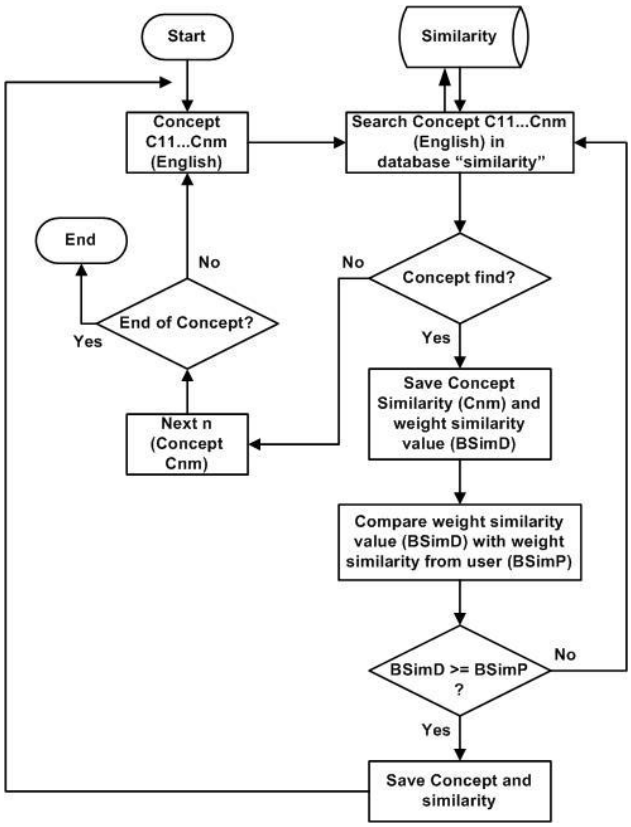

Fig. 3. Steps to searching concept similarity.

Once the concept similarity is found and stored along with its similarity weights (BSimD), then carried the weight of similarity (BSimD) ratio with similarity weight is entered by the user (BSimP). If BSimD value greater than or equal to BSimP value (BSimD>=BSimP), then the equivalence of concept and similarity weights (BSimD) is stored as the final result in the discovery stage concept similarity (similarity weight range is 0 to 1).

Then continue the search for the same concepts in the "similarity" database, because it could be the concept similarity has more than one stored in the database.

Concept Similarity with reference to the weight of similarity is in the "similarity" database with the similarity weights entered by the user can be created mathematical notation as in equation (2) refers to the equation (1).

$$Q \rightarrow \{C1 \ldots Cn\}, \text{ when } n>1 \quad (1)$$

Description:
Q=First Query
C=Concept
n=Concept index

$$\begin{aligned} &\text{WordSim } C11 \rightarrow \{C11 \ldots C1m\}, \text{ when } m>1 \\ &\text{WordSim } C21 \rightarrow \{C21 \ldots C2m\}, \text{ when } m>1 \end{aligned} \quad (2)$$

.
.
WordSim C*n*1 → {C*n*1 . . . C*nm*}, when *n*>1 and *m*>1
Description:
WordSim=concept similarity
C=concept
m =array variable for concept

The following is the first algorithm that is used to perform a concept similarity search.

Algorithm 1. Searching for Concept Similarity
Input: *K*1 . . . *Kn*

Process:
1 JumK←amount of concept from query extraction;
2 BSimP← user *similarity* value;
3 *n*←1;
4 *m*←1;
5 For *n*←1 to JumK do
6 Begin
7 Read for concept *K*(*n*);
8 Find *K*(*n*) in field "Kata1" from "similarity" database
9 q ← 1
10 While Kata1=*K*(*n*) do
11 *C*(*n*, *m*+*q*) ← Kata2 in "similarity" database;
12 If BSimD (*n*, *m*+*q*)>=BSimP then
save (*K*(*n*); *C*(*n*, *m*+*q*));
13 Else goto 16;
14 JumWSi ←*q*;
15 *q*←*q*+1;
16 Find next *C*(*n*, *m*) in field "Kata1" from database
17 Until *C* (*n*, *m*) not found in database
18 End
19 Next *n*;
20 End.

Output: C_{11} ... C_{nm}

Description of Algorithm:
JumK=The number of concepts extracted query
BSimP=similarity weights are entered by the user
BSimD=similarity weights obtained during the search in the "similarity" database
K=concept from extraction query results
C=concept similarity that found from "similarity" database
JumWSim=number of concept similarity generated after searching the "similarity" database
n, *m*, *q*=array index

In the algorithm as input is the concept resulting from the extraction query process. Then searched whether the concept were read in the column "Kata1" in "similarity" database is found. If found, the concept store in the variable C, then compare the similarity weight is in the database with similarity weights desired by the user, if the weight of the database similarity greater than or equal to the weight of the desired user similarity, then the concept has found the appropriate term.

IV. Searching Concept Similarity Illustration

Illustration of the concept similarity search algorithm was developed is shown in Fig. 4.

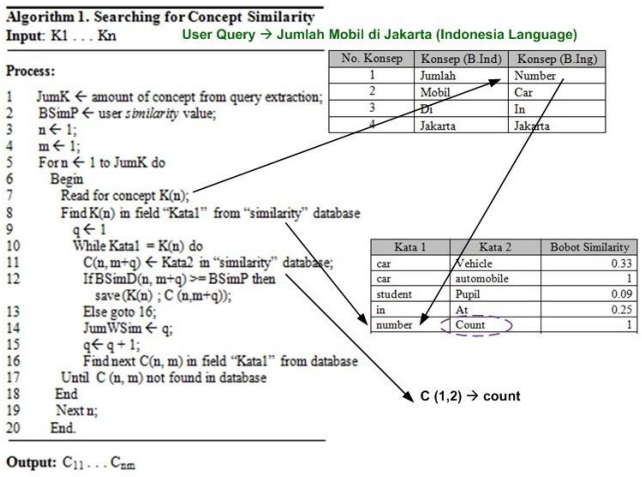

Fig. 4. Algorithm searching for concept similarity illustration.

Algorithm in this article used sentences in Indonesian Language (Bahasa). Input for the algorithm was sentence that user used, and in this illustration, user query as an input was sentence "Jumlah Mobil di Jakarta" (in English is defined "Numbers of Car in Jakarta").

In the illustration Fig. 4, Step 7 is reading concepts generated in the extraction query process. The following is an illustrative example. Step 7 is to read the concept-1="Jumlah" (in English is defined "Number of"), and it was converted into English becomes "Number". In step 8 searched the word "Number" on "similarity" database. At step 10 and 11 if found the word "Number", then read the contents of the column "kata2" in the database. Found the word "Count", so that the "Count" is stored in the variable *C* (1, 2)="Count".

Step 12 (shown in Fig. 4) is to compare the similarity weights were obtained from the database with similarity weight entered by the user, if they match the conditions, then *C* (1, 2)="Count" is stored as the concept similarity, and the number of concept similarity (JumWSim) is 1 then steps will be repeated to step 10, and so on until all the matching concept is found.

V. Conclusion

Information search on the Internet through search engines using keyword, and to provide more relevant search results to the user's wishes, it would require an expansion of the query, by using the concept of query rewriting.

The approach that is developed in this article is the approach that is used to find the concept similarity. It is a process that must be done before the query rewriting process.

In further research, the approach to find similarity of this concept will be used to carry out the whole process of query rewriting. Concept similarity that is found is used to rearrange the query into other queries that are commensurate with the meaning of key words from the initial query entered by the user at the time of the search process on the Internet using search engines.A conclusion section is usually required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

Acknowledgment

This work is partially supported by Gunadarma University, the Doctoral Program of Information Technology at Gunadarma University Jakarta, and YAI Jakarta Indonesia.

References

[1] R. Mandala, "Evaluasi efektifitas metode machine-learning pada search-engine," in *Proc. Seminar Nasional Aplikasi Teknologi Informasi*, Yogyakarta, 2006.
[2] P. W. Handayani, I. M. Wiryana, and J. T. Milde, "Mesin pencari berbasiskan semantik untuk bahasa indonesia," *Journal Sistem Informasi*, vol. 4, no. 2, 2008.
[3] A. Shiri and C. Revie, "Query expansion behavior within a thesaurus-enhanced search environment: A user-centered evaluation," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 4, 2006.
[4] H. Imran and A. Sharan, "Thesaurus and query expansion," *International Journal of Computer Science & Information Technolog*., vol. 1, no. 2, 2009.
[5] Z. Kunpeng, W. Xiaolong, and L. Yuanchao, "A New query expansion method based on query logs mining," *International Journal of Asian Language Processing*, vol. 19, no. 1, 2009.
[6] I W. S. Wicaksana, *Survei dan Evaluasi Metode Pengembangan Ontology*, Universitas Gunadarma, Indonesia, 2004.
[7] L. Y. Banowosari and I W. S. Wicaksana, *Pemeliharaan Common Ontology Pada P2P Dengan Voting dan Representasi*, Yogyakarta, 2005.
[8] B. Alexander and H. Graeme, "Evaluating WordNet-based measures of lexical semantic relatedness," *Journal Computational Linguistics*, vol. 32, pp. 13-47, 2006.
[9] S. Giriprasad, H. Emily, P. Lori, S. K. Vijay, "Identifying word relations in software: A comparative study of semantic similarity tools," in *Proc. the 16th IEEE International Conference on Program Comprehension*, 2008, pp. 123-132.

Detty Purnamasari was born in Balikpapan, East Kalimantan, Indonesia on May 18, 1981. She has the educational background in information system for bachelor degree and accounting information system for master degree. She finished her study at Gunadarma University Jakarta Indonesia.

She is a secretary and a lecturer at Gunadarma University, and since 2013 she's finished her PhD degree in information technology at Gunadarma University Jakarta Indonesia.

Her research topic about information system and the main focus is how to collect datas in internet, web semantic, query rewriting. Besides, she is interest in the research about economic and tries to combine information system and economic sciences.

Lily Wulandari was born in Jakarta, Indonesia on March 6, 1969. She completed the doctoral program of information technology at Gunadarma University, Jakarta, Indonesia in 2009, got the graduate program, magister of information system management at Gunadarma University, Jakarta Indonesia in 1998 and the bachelor program in information system at

Gunadarma University, Jakarta, Indonesia in 1994.

She is a lecturer at the Gunadarma University. She has been involved in several projects in the information system of government and the private sector in Indonesia. She produced several scientific papers which are mainly themed information interoperability.

Ahmad Muhammad Thantawi was born in Sampit Central Kalimantan-Indonesia on July 11, 1974. He completed the doctoral program of information technology at Gunadarma University, Jakarta, Indonesia in 2014, got the graduate Program and magister of information system management at Bina Nusantara University, Jakarta Indonesia in 2006 and the bachelor program in informatic engineering at Gunadarma University, Jakarta, Indonesia in 1997.

Since 2003, he worked as a lecturer at the Department of Informatics Engineering, Faculty of Engineering, University of Persada Indonesia YAI. Currently, he works as a director in University of Persada Indonesia YAI in Jakarta Indonesia. His current research is in the field of databases, especially semantic and query rewriting.

I Wayan Simri Wicaksana was born at Surabaya, East Java, Indonesia on June 11, 1964. He graduated with the bachelor degree in physic from University of Indonesia at 1988, then he continued the master program in Computer Integrated Manufacturing Swinburne University, Melbourne Australia and finished the doctoral program in informatics from University of Dijon in France and Gunadarma University in Jakarta Infonesia.

He is a professor in information technology. His main research areas are in database, semantic web, interoperability, and ontology.