Wiki-MetaSemantik: A Wikipedia-derived Query Expansion Approach based on Network Properties

Diyah Puspitaningrum
and Gries Yulianti
Dept. of Computer Science
University of Bengkulu
Bengkulu, Indonesia
Emails: diyahpuspitaningrum@gmail.com,
gries.cs@unib.ac.id

I.S.W.B. Prasetya
Dept. of Information and
Computing Sciences
Utrecht University
Utrecht, Netherlands
Email: s.w.b.prasetya@cs.uu.nl

***Abstract*—This paper discusses the use of Wikipedia for building semantic ontologies to do Query Expansion (QE) in order to improve the search results of search engines. In this technique, selecting related Wikipedia concepts becomes important. We propose the use of network properties (degree, closeness, and pageRank) to build an ontology graph of user query concepts which is derived directly from Wikipedia structures. The resulting expansion system is called Wiki-MetaSemantik. We tested this system against other online thesauruses and ontology based QE in both individual and meta-search engines setups. Despite that our system has to build a Wikipedia ontology graph in order to do its work, the technique turns out to work very fast (1:281) compared to another ontology QE baseline (Wikipedia Persian ontology QE). It has thus the potential to be utilized online. Furthermore, it shows significant improvement in accuracy. Wiki-MetaSemantik also shows better performance in a meta-search engine (MSE) set up rather than in an individual search engine set up.**

***Index Terms*—Wiki-MetaSemantik, meta-search engine, query expansion, semantik, wikipedia**

I. Introduction

Wikipedia is the largest human-built knowledge repository currently in existence, available in over 250 languages, with characteristics such as: it is not limited in scale, includes dense link structures, URL-based word sense disambiguation, and brief anchor [14][8]. Wikipedia's structures consist of: articles, disambiguation pages, redirects, hyperlinks, category structure, templates and infoboxes, discussion pages, and edit histories. An article in Wikipedia describes a single concept. Each article's title resembles terms in a conventional thesaurus. Terms of similar meaning are linked to an article using redirect. The disambiguation pages allow users to select an article. Hyperlinks in an article express relationships to other articles. All of these structures form a various number of concepts of human knowledge that can be exploited as a tool to build a semantic ontology of any concept.

Since currently available Wikipedia ontologies only map small parts of human knowledge concepts, in this paper, we propose a new approach of query expansion (QE), called Wiki-MetaSemantik, to build an ontology automatically from Wikipedia. Wiki-MetaSemantik exploits all the benefits of Wikipedia structures and alleviates topic drift that otherwise often appears when using synonyms of online thesauruses such as WikiSynonyms, WordNet, and Moby. Existing Pseudo-Relevance Feedback (PRF) query expansion methods are still suffer from several drawbacks such as query-topic drift [9][11] and inefficiency [22]. Al-Shboul and Myaeng [1] proposed a technique to alleviate topic drift caused by words ambiguity and synonymous uses of words by utilizing semantic annotations in Wikipedia pages, and enrich queries with context disambiguating phrases. Also, in order to avoid expansion of mistranslated words, a query expansion method using link texts of a Wikipedia page has been proposed [10]. Furthermore, since not all hyperlinks are helpful for QE task (e.g. a link text "French" inside a page titled "baseball" is not very helpful to expand a query on the latter), Farhoodi et al. [4] proposed a QE method using ontology derived from Wikipedia, or the Wikipedia Persian ontology method for short. They used weights to capture relationships between Wikipedia structures.

In our approach, Wiki-MetaSemantik, we alleviate topic drift by creating an ontology graph of concept (or topic) on-the-fly from many related Wikipedia pages, and then choose important terms based on network properties (degree, closeness, and pageRank). The utilization of network properties involve less computation than other QE approach based on ontology such as [4]. In selecting QE, terms distributions and structures of Wikipedia pages are taken into account. Since combining multiple datasets can lead into better accuracy [15], viz. a meta-search engine [19], we propose the use of meta-pseudo relevance feedback (meta-PRF) for automatic judgement purpose as in [18][19]. Our experiments show improvement in IR performance when using Wiki-MetaSemantik.

The main contribution of this work is the approach behind Wiki-MetaSemantik, which enables fast relevant building of Wikipedia ontology from queries of any concept. Despite its modesty, we believe the algorithm to be potential for running on-the-fly thus it can be embedable in either a meta-search or individual search engine. By building an ontology online, our algorithm does not really depend on how extensive, or limited, related concepts are documented in Wikipedia. Hence, the algorithm is more flexible in capturing the semantic of any

user query.

The remainder of this paper is organized as follows. Section 2 discusses related research on QE. Section 3 describes our proposed methods for using ontology-derived from Wikipedia as meta-pseudo relevant documents. Experimental results are reported in Section 4. We summarizes the outcomes, possible limitations, and the future work directions in Section 5.

II. Related Work

Hsu et al. [5] shows the use of WordNet in query expansion. Performance of queries expanded by WordNet outperforms that of queries without expansion, and queries expanded with a single resource. Semantic graphs are commonly used to model word senses and are usually built using thesauri or lexical databases such as WordNet [20]. Approaches such as PageRank, HITS or node similarity can be used to second alternative queries [12][21]. Bruce et al. [2] uses Wikipedia and its hyperlink structures to find related terms for reformulating a query using link probability weighting and link based measure.

The work by Farhoodi et al. built query expansion for Persian ontology [4]. To improve the results of retrievals, they proposed to exploit the following: 1) the relation between title and keywords, 2) the relation between the title and the concepts in article's text, and 3) the relation between the title and the concepts in 'See also' links. These relations are given the weights of 0.6, 0.5, and 0.7 respectively. Most of the results have higher precision when the query expansion is implemented but the precision may fall depending on the quality of Wikipedia pages and the links in these pages.

Our query expansion method takes advantages of the above mentioned work [4], by reusing the set up weights of Wikipedia relations. However, the method itself works in very differently. Assume we have a graph/network of a small world (e.g. a set of selected Wikipedia concepts). A user query becomes the root node and structures such as titles, keywords, text, 'See also', and 'Category' become leafs or nodes in an ontology graph, up to a certain number of hops of Wikipedia pages. Each node has links to other Wikipedia pages, which become the edges in the graph. QE terms are generated using a carefully set up weights and combinations of graph-based measures (degree, closeness, and pageRank).

Graph centrality measures are used to determine how important a node is in a network/graph [16]. In degree centrality, a node is important if it has many edges connected to and from it. Network degree is the maximum degree centrality over a network's nodes. In closeness centrality, a node is important if it is "close" to all other nodes in the network, in terms of the sum of the shortest paths to all other nodes [6]. In PageRank, a term is as valuable as other terms that link to it. According to Page et al. [17], a typical analogy for this is that a link from one page to another essentially can be seen as a vote being cast by one page onto another. In our query expansion case, a node is voted by the number of its backlinks (the links from other nodes to that node).

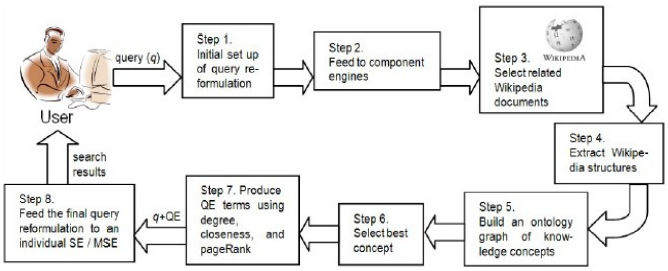

Fig. 1. Overview of Wiki-MetaSemantik search engine system

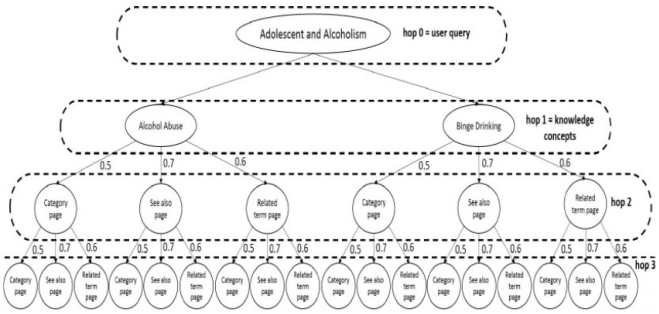

Fig. 2. A typical or modest example of an ontology graph from query "*adolescent and alcoholism*" (truncated from original due to space)

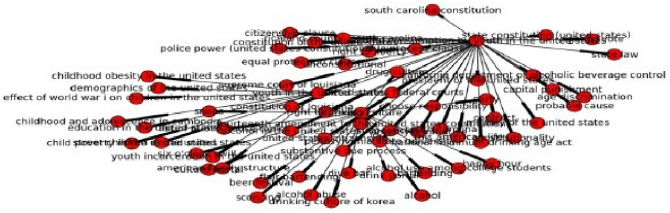

Fig. 3. First Hop of "*alcohol consumption by youth in the united states*" concept

III. Wiki-MetaSemantik: The Process

When Wiki-MetaSemantik receives a query (Fig. 1), it will be processed through the following eight steps.

Step 1: Initial set up of query reformulation. For each query, we add it with "wikipedia" word. For example query "adolescent and alcoholism" has initial query format of "adolescent and alcoholism wikipedia".

Step 2: Feed the initial query reformulation to component engines. Feed the initial query to each component engines of a meta-search engine, or to an individual one.

Step 3: Select related Wikipedia documents. Choose only top-documents originated from any Wikipedia domains.

Step 4: Extract Wikipedia structures. Visit each Wikipedia pages and extract related keywords from the pages by exploiting all Wikipedia structures: Title, 'Category', 'See also', and 'Related terms'. 'Related terms' denotes all terms in the Wikipedia passages that marked by blue fonts (hyperlinks, or assumed as relevant regarding to the corresponding Title). Each initial Wikipedia pages (hop 1), as a result of Step 3, is a candidate of knowledge concepts

of the user query. For instance the user query "*adolescent and alcoholism*" results 8 candidates of knowledge concepts: 1) Alcoholism (https://en.wikipedia.org/wiki/alcoholism), 2) Alcoholism in adolescence (https://en.wikipedia.org/wiki/alcoholism_in_adolescence), 3) Alcohol abuse (https://en.wikipedia.org/wiki/alcohol_abuse), 4) Binge drinking (https://en.wikipedia.org/wiki/binge_drinking), 5) Alcohol consumption by youth in the United States-Wikipedia (https://en.wikipedia.org/wiki/alcohol_consumption_by_youth_in_the_united_states), 6) Alcoholism in family systems (https://en.wikipedia.org/wiki/alcoholism_in_family_systems), 7) Substance abuse (https://en.wikipedia.org/wiki/substance_abuse), 8) Alcohol and health (https://en.wikipedia.org/wiki/alcohol_and_health).

Step 5: Build an ontology graph of knowledge concepts. Do Step 4 to every existing Wikipedia hyperlinks until third hop, and then create an ontology graph (Fig. 2). Fig. 3 shows an ontology graph generated from a part of nodes of user query "*adolescent and alcoholism*" which represented by "*alcohol consumption by youth in the united states*" concept.

Step 6: Select best concept. Isolate the ontology graph into small graphs, starts by taking candidates of knowledge concepts as roots of each those small graphs and then respectively go down to all descendants (or all related nodes). Choose one of the isolated graphs with the highest degree as the best concept. Degree of graph shows a graph importance.

Step 7: Produce QE terms using degree, closeness, and pageRank. For all nodes in the best concept, compute their degree, closeness, and pageRank scores and then create lists of nodes in decreasing order sorted separately by their degree, closeness, and pageRank scores. After that, find sets of nodes (or keywords) that appear in degree, closeness, and pageRank lists. Different ways of intersection the lists will produce different keywords and different ordering. The following three ways to construct intersections are used: 1) Intersection_set#1: Find 100-top nodes from the degree-ordered list that also appear in the closeness-ordered list and the pageRank-ordered list; 2) Intersection_set#2: Find 100-top nodes from the closeness-ordered list that also appear in the degree-ordered list and the pageRank-ordered list; 3) Intersection_set#3: Find 100-top nodes from pageRank-ordered list that also appear in closeness-ordered list and degree-ordered list.

The three intersection lists are then combined into one list using the Borda Count voting technique [3]. Post-processing is then applied by filtering the resulting list to take only terms that are neither in user query nor stopwords. Final QE terms, or just QE terms for simplicity, are taken from the remaining Borda Count list for top-*2* or top-*3* terms depending on user needs.

Step 8: Feed the final query reformulation to an individual search engine or to a meta-search engine. The final query reformulation is defined by: user query + QE terms. Fetch them to an individual or to a meta-search engine (MSE) along with a method to get search results. The weights are input parameters to create an MSE.

IV. Experimental Evaluation

TABLE I
The Basic Multi Domain Queries [18][13]

Two Terms	Three Terms
database overlap	comparative education methodology
multilingual OPACs	java applet programming
programming algorithm	indexing digital libraries
roadmap plan	geographical stroke incidence
adolescent alcoholism	culturally responsive teaching

A. Test Data

Table I shows a set of basic queries we will use in our experiments. They are expanded to 30 multi domain queries using combinations of operator 'AND/OR'. We use four baseline methods against which Wiki-MetaSemantik will be compared to. Method 1 to 3 use synonyms from online thesauruses (WordNet, Moby Thesaurus, or WikiSynonyms respectively), method 4 use QE using the Wikipedia Persian ontology method [4].

We did all of our experiments on both individual search engines and meta-search engines from 5 popular search engines: Google, Bing, Lycos, Ask, and Exalead. For the meta-search engines, we use 10 combinations of three component search engines: (Google-Lycos-Bing), (Google-Lycos-Ask), (Google-Lycos-Exalead), (Google-Bing-Ask), (Google-Bing-Exalead), (Google-Ask-Exalead), (Lycos-Bing-Ask), (Lycos-Bing-Exalead), (Lycos-Ask-Exalead), and (Bing-Ask-Exalead).

All experiments were run on a machine with 2 GHz CPU, 2 GB memory, and 7.2 Mbps network connection. Wiki-MetaSemantik is developed using Python version 2.7.9, running on Windows 10.

B. Measurements

1) Baselines: As mentioned before, four baselines are used in our evaluation. The WordNet and the WikiSynonyms QE methods work by searching over top-*k* synonyms in either WordNet or WikiSynonyms. The Moby thesaurus QE method works by searching over *k* random synonyms in the Moby Thesaurus. Both the WordNet results and the WikiSynonyms results are sorted in a decreasing order by synonymity, whereas the Moby results are ordered in equal weight of synonymity.

In the Wikipedia Persian Ontology QE, the concepts in the query are mapped on to the ontology graph, based on Wikipedia relationship. The ontology is then used to expand user queries and submitted to the search engine to get the search results. Wiki-MetaSemantik is simpler than the Persian Wikipedia Ontology QE because it captures the most significant terms from a concept graph using network properties only.

2) Ranking Suggestions: For all QE methods used in this paper, the postprocessings found QE terms are post-processed as follows: ignore the AND/OR operators and delete stopwords from user query. Then, generate QE terms by taking related keywords per user query term and in the FCFS order.

MetaPRF: Generate $\Re$

Input:

user_query : a user query

#QE_terms : number of terms in QE

QE_degree($g_{concept}(qc_x)$), QE_closeness($g_{concept}(qc_x)$), QE_pageRank($g_{concept}(qc_x)$), QE_WordNet($g_{concept}(qc_x)$), QE_WikiSynonyms($g_{concept}(qc_x)$), QE_mobyThesaurus($g_{concept}(qc_x)$) : list of candidates of QE terms in decrease order as a result of computing {degree|closeness|pageRank|WordNet|WikiSynonyms |mobyThesaurus} from a Wikipedia graph derived from a query concept qc_x given user_query

Output:

$\Re$: a web search results list of a meta-search engine (MSE)

1. Define number of query expansion terms, m = #QE_terms .
2. DO query rewriting:
   Expanded_query = user_query + top_m {QE_degree($g_{concept}(qc_x)$)| QE_closeness($g_{concept}(qc_x)$)| QE_pageRank($g_{concept}(qc_x)$) | QE_WordNet($g_{concept}(qc_x)$)| QE_WikiSynonyms($g_{concept}(qc_x)$) | QE_mobyThesaurus($g_{concept}(qc_x)$) }
3. Fetch each expanded query in Step 2 to each component engines separately.
4. FOR each component engines SE_i, i=1,...,n where n is total number of component engines:
   4.1. Set up confidence values for component engines.
   // e.g. $SE\_conf_i = 30$ for Google, $SE\_conf_i = 25$ for Lycos, $SE\_conf_i = 20$ for Bing, $SE\_conf_i = 15$ for Ask.com, $SE\_conf_i = 10$ for Exalead.
   4.2. Set up weights for 6 knowledge sets.
   // e.g. w_degree = 30, w_closeness = 20, w_pageRank = 20, w_WordNet = 10, w_WikiSynonyms = 10, w_mobyThesaurus = 10
   4.3. Set up weight values of each component engines:
   $w\_SE\_degree = w\_degree * SE\_conf_i / 100$
   $w\_SE\_closeness = w\_closeness * SE\_conf_i / 100$
   $w\_SE\_pageRank = w\_pageRank * SE\_conf_i / 100$
   $w\_SE\_WordNet = w\_WordNet * SE\_conf_i / 100$
   $w\_SE\_WikiSynonyms = w\_WikiSynonyms * SE\_conf_i / 100$
   $w\_SE\_mobyThesaurus = w\_mobyThesaurus * SE\_conf_i / 100$
5. Create an MSE using top-200 search results from component engines. (See [19]).

Fig. 4. The Meta-PRF algorithm: a gold standard algorithm using data fusion

TABLE II
AVERAGE RUNNING TIME PER QUERY IN INDIVIDUAL SEARCH ENGINES (IN SECONDS). WIKI-METASEMANTIK IS SET WITH THE DEGREE/CLOSENESS/PAGERANK PARAMETERS OF (20,30,20).

QE Method	#QE_terms=2	#QE_terms=3
Wiki-MetaSemantik, #query_terms=2	0.473	0.328
Persian Ontology, #query_terms=2	33.283	36.160
Wiki-MetaSemantik, #query_terms=3	0.249	0.327
Persian Ontology, #query_terms=3	70.030	53.749

For example: for a two terms basic query "*adolescent and alcoholism*", if number of QE terms=2 then we pick one synonym from "*adolescent*" and "*alcoholism*" respectively; if number of QE terms=3 then we pick two synonyms from "*adolescent*" and one from "*alcoholism*". For a three terms basic query "*Java and applet and programming*", if number of QE terms=2 then pick one synonym from "*Java*" and one synonym from "*applet*" only; if number of QE terms=3 then pick one synonym from each word.

3) Evaluation Utility: Automatic Judgement: Instead of asking the user to identify relevant documents, we simply assume that the top-ranked documents are relevant (*pseudo-relevance feedback*). To do automatic relevance judgement, we compare search results over a meta-search engine or an individual search engine against a gold standard, viz. the top-k search results of the Meta-PRF algorithm, k={3,5,10,20,50} with ten QE terms. We choose the top-*200* retrieved documents of each component engine to be merged as an MSE search results because they are decreasingly ordered by relevance and that they are the most probable by-user viewed documents. We define the number of QE terms=10 under the assumption that the top-*10* terms of each knowledge sources (WordNet, WikiSynonyms, Moby thesaurus, degree, closeness, pageRank) are higly related terms and they capture well the concept and semantic of user queries. Following (Fig. 4) is an algorithm for creating the gold standard, viz. the pseudo-relevant dataset. The meta-PRF algorithm takes advantage of data fusion: it combines advantages of each component engine. The MSE algorithm viz. Weighted Borda Fuse (or WBF) is as in [19][7]. As inputs we take only synonyms of user query from online thesauruses (WordNet, WikiSynonyms, Moby Thesaurus), as well as synonyms and related terms from Wikipedia by computing ontology graph properties (degree, closeness, pageRank) as in Step 7 in Section 3.

As evaluation criteria we use precision, success and runtime, denoted by P@x, S@x and time. P@x denotes precision of the x highest ranked documents with $x\in\{5,10,20,50\}$, and is defined as the average percentage of the first x retrieved documents that is relevant with the gold standard, averaged over all documents. Similarly, S@x denotes success of the x highest ranked documents.

4) Evaluation Utility: Human Judgement: For the ground truth, we use Cohen's kappa coefficient to measure the reliability of scoring diagnosis by two human judges. For each query, we take top-*10* retrieved documents from each QE methods to be scored either 0, 1, or 2 where 0 = ("*not relevant*"), 1 = ("*partially relevant*"), and 2 = ("*relevant*"). Then the Cohen's kappa coefficient measures agreement between judges on the same objects and subtracting out agreement due to chance. The kappa coefficient, or κ, has value around [0,1]. The closer the κ coefficient to 1 the more agree the two parties. Once κ coefficient shows strong agreement of 2 judges, quality of query expansion system is measured using Normalized DCG (NDCG).

C. Results and Discussion

Table II shows the scalability of Wiki-MetaSemantik in terms of its running time, to start once an ontology graph is created. Persian Wikipedia Ontology QE is many times slower due to it involves vector processing with four times ontology matrix multiplication where the maximal index of the query vector and ontology matrix is equal to the number of nodes in the ontology graph. On the other hand, Wiki-MetaSemantik works very fast because it computes degree, closeness, and pageRank at once, and is thus time efficient. From Table II, the ratio of running time between Wiki-MetaSemantik vs Persian Wikipedia Ontology QE is ranging between 1:70 and 1:281 when the number of QE terms=2 and 1:110 and 1:164 when the number of QE terms=3, with the first two is for number of user query=2 and the latter two is for number of user query=3. This suggests that an online implementation of WikiMetaSemantik would be very attractive.

Parameters tuning in Wiki-MetaSemantik should be treated carefully. We did experiments with different weights of graph properties (degree-closeness-PageRank) viz. (30-20-20), (20-30-20), (20-20-30), and found that (20-30-20) is the best parameters set up (see Fig. 5). This shows the importance of *closeness* weight parameter, followed by *degree* weight and

Fig. 5. P@3 of Wiki-MetaSemantik vs Persian Ontology with number of QE terms = 2 (left) and 3 (right).

Fig. 6. Precisions of Wiki-MetaSemantik (degree-closeness-pageRank)=(20-30-20)). Left to right: without QE, number of QE terms=2, number of QE terms=3. All is running on MSE (Lycos-Bing-Exalead).

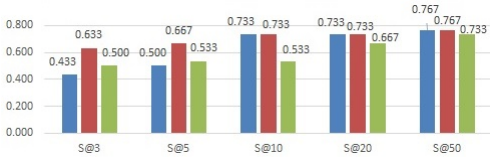

Fig. 7. Successes of Wiki-MetaSemantik (degree-closeness-pageRank)=(20-30-20)). Left to right: without QE, number of QE terms=2, number of QE terms=3. All is running on MSE (Lycos-Bing-Exalead).

pageRank weight. Giving more weight to a graph properties when creating an MSE means we trust more on QE term candidates from its graph properties. Giving more weight to pageRank turns out to be not very helpful (Fig. 5(c)) since the ontology graph actually resembles more to a tree than a cyclomatic graph. Furthermore, the higher the closeness score to a knowledge concept, the more important the node (or the QE term candidate) is.

Using the same set up of graph properties, Fig. 6 and Fig. 7 show that Wiki-MetaSemantik improves the relevance of retrieved documents at top-*5* documents; the relevance decreases after that. This means that Wiki-MetaSemantik is working well on the locations of the top most viewed documents by search engine users. By Fig. 6 the improvement ratio scores for precisions are: (1:1.62:1.62) and (1:1.18:1.09) which are from P@3 and P@5, and by Fig. 7 the improvement ratio scores for successes are: (1:1.46:1.16) and (1:1.33:1.07) which are from S@3 and S@5. This is a significant improvement, especially in the top-*3* retrieved documents, that would significantly help users finding relevant documents.

Table III shows an example of gold standard QE terms against QE terms from several QE methods. The m=10 in Fig. 4 means the gold standard has 10 QE terms taken from top-10 of BC list of Intersection_set#1, Intersection_set#2, and Intersection_set#3 (see Section 3 Step 7). Table III shows the 10 QE terms in gold standard are qualified because it shows highly related terms with user query.

Table IV shows some examples of QE of our total 30 multi-domain queries. All the terms in the figure are those suggested by each QE methods. It shows that Wiki-MetaSemantik is very good in capturing semantic relatedness. It retrieves highly related QE terms because Wiki-MetaSemantik uses only nodes from best ontology (or best knowledge concept), and filter the nodes using graph properties of (degree, closeness, and pageRank) to produce qualified QE terms.

Furthermore, we prove the quality of our Wiki-MetaSemantik search results by random sampling multi domain queries and ask 2 judges for relevance. We found that the κ coefficient is ranged from 0.512 to 1, which shows fair to good agreement between the judges. Table V shows average of NDCG scores of two human judges at top-k search results documents (k=$\{3,5,7,10\}$) for Wiki-MetaSemantik on number of QE terms=$\{2,3\}$. From the figure, Wiki-MetaSemantik with number of QE terms=2, even shows an ideal ranking until top-*10* of retrieved documents. More QE terms can lead to bias against user query.

In general, Wiki-MetaSemantik shows significant improvement of performance (Fig. 6 and Fig. 7). Wiki- MetaSemantik works better while implemented in an MSE rather than in an individual search engine (Fig. 5) because an MSE combines all advantages of its component engines. Fine tuning of graph properties' weights influences its performance with their order of importance as follows: closeness>degree>pageRank. The structure of Wikipedia derived ontology graph influences the pageRank performance due to pageRank being good in a cyclomatic graph rather than a tree graph. Allowing indirect links of Wikipedia pages, the minimum length of path (*closeness*) from user query concept is more important than the number of outlinks a node has (*degree*). This simplicity as well as an idea of generating an ontology graph on-the-fly make it suitable for multi domain queries thus we do not have to depend on limited ontologies available in Wikipedia. Also since it works very fast against other Wikipedia query expansion baselines (Table II) thus Wiki-MetaSemantik is potential to be implemented online.

TABLE III
#QE_TERMS = {2,10} FOR QUERY "ADOLESCENT AND ALCOHOLISM". META-PRF QE TERMS WERE GENERATED FROM LYCOS-BING-EXALEAD

WIKI-META SEMANTIK	PERSIAN ONTOLOGY	WIKI SYNONYMS	WORDNET	MOBY
alcoholic	pejorative	adolescence	stripling	juvenal
beverage	alcohol	alcoholic	alcohol	drug
QUERY REWRITING (META-PRF, #QE_TERMS=10):				
DEGREE:	(adolescent and alcoholism) dmoz stereotype public health disability-adjusted life year cocaine addiction ethanol			
CLOSENESS:	(adolescent and alcoholism) dmoz stereotype cocaine addiction public health disability-adjusted life year ethanol			
PAGERANK:	(adolescent and alcoholism) self-medication alcohol withdrawal syndrome addictive personality benzodiazepine physical dependence addiction			

TABLE IV
EXAMPLES OF QE TERMS. BAD TERMS ARE IN UNDERLINES.

WIKI-META SEMANTIK	PERSIAN ONTOLOGY	WIKI SYNONYMS	WORDNET	MOBY
#QE_terms = 2				
user query = "*adolescent and alcoholism*"				
alcoholic	pejorative	adolescence	stripling	juvenal
beverage	alcohol	alcoholic	alcohol	drug
user query = "*programming or algorithm*"				
coding	adaptive-	computer	scheduling	mapping
theory	additive entropy	algorithmics	algorithmic	formula
user query = "*geographical or stroke or incidence*"				
public	hypercholesterol	geography	geographic	science
health	emia cincinnati	cerebrovascular	shot	stress
#QE_terms = 3				
user query = "*adolescent and alcoholism*"				
alcoholic	pejorative	adolescence	stripling	young
beverage	alcohol	teenage	teenage	teener
psychological	cardiovascular	alcoholic	alcohol	drug
user query = "*programming or algorithm*"				
coding	adaptive-	computer	scheduling	document
theory	additive entropy	software	programming	rule
information	binary	algorithmics	algorithmic	formula
user query = "*geographical or stroke or incidence*"				
public	hypercholesterol	geography	geographic	science
health	emia cincinnati	cerebrovascular	shot	stress
methampine	scale	angle	relative	routiness

V. CONCLUSION

We have proposed Wiki-MetaSemantik, a query expansion technique using an ontology graph derived from Wikipedia that captures semantic relatedness very well. The results show that it can improve relevance scores of retrieval system as well as its efficiency.

A possible limitation of Wiki-MetaSemantik is the quality of QE terms it produced, which may fall depending on the quality of Wikipedia pages and the links in these pages. Therefore for future work, we suggest the tight integration of Wiki-MetaSemantik with online dictionary and thesauruses to enrich vocabularies so they can help in improving the success of ontology-based query expansion terms over the plain (only synonyms) user query terms to cover up lack of knowledge due to inexistence of Wikipedia pages on certain topics.

TABLE V
WIKI-METASEMANTIK PERFORMANCE (HUMAN JUDGEMENT) OF USER QUERY "ADOLESCENT AND ALCOHOLISM" WITH SET UP FOR GRAPH PROPERTIES (DEGREE-CLOSENESS-PAGERANK)=(20-30-20)

#QE terms	$NDCG_3$	$NDCG_5$	$NDCG_7$	$NDCG_{10}$
2	1.00	1.00	1.00	1.00
3	0.81	0.73	0.83	0.90

REFERENCES

[1] B. Al-Shboul and S.-H. Myaeng, "Query phrase expansion using wikipedia in patent class search," in *AIRS 2011, LNCS 7097*, Berlin: Springer-Verlag, 2011, pp.115-126.

[2] C. Bruce et al., "Query expansion powered by wikipedia hyperlinks," in *AI 2012, LNCS 7691*, M. Thielscher and D. Zhang, Eds. Berlin: Springer-Verlag, 2012, pp.421-432.

[3] C. Dwork et al., "Rank aggregation methods for the web," in *Proc. of the ACM International Conference on World Wide Web (WWW)*, pp.613-622, 2001.

[4] M. Farhoodi et al., "Query expansion using persian ontology derived from wikipedia," *World Applied Sciences Journal*, vol. 7 no. 4, pp.410-417, 2009.

[5] M.-H. Hsu et al., "Query expansion with conceptnet and wordnet: an intrinsic comparison," in *AIRS 2006, LNCS 4182*, H.T. Ng et al., Eds. Singapore: Springer, 2006, pp.1-13.

[6] I. Hulpus et al., "Unsupervised graph-based topic labelling using dbpedia," *6th ACM Int'l Conf. on Web Search and Data Mining (WSDM'13)*, pp.465-474, 2013.

[7] H. Jadidoleslamy, "Search result merging and ranking strategies in meta-search engines: a survey," *J. Computer Science*, vol. 9 no. 4, pp.239-251, 2012.

[8] A.M. Jimenez, "Using wikipedia to improve web search discovery," Ph.D. dissertation, Science and Engineering Faculty, School of Electrical Engineering and Computer Science, Queensland University of Technology, Brisbane, Australia, 2012.

[9] H. Lang et al., "Improved latent concept expansion using hierarchical markov random fields," *Proc. of CIKM 2010*, pp.249-258, 2010.

[10] T. Lin and S. Wu, "Query expansion via wikipedia link," *Int'l Conf. on Information Technology and Industrial Application*, Queensland, Australia, 2008.

[11] Y. Lv and C. Zhai, "Adaptive relevance feedback in information retrieval," *Proc. of CIKM 2009*, pp.255-264, 2009.

[12] C. Makris et al., "Web query disambiguation using pagerank," *J. Am. Soc. Inf. Sci. Tech.*, vol. 63 no. 8, pp.1581-1592, 2012.

[13] K.A-E-F. Mohamed, "Merging multiple search results approach for meta-search engines," Ph.D. dissertation, Graduate Faculty of School of Information Sciences, University of Pittsburgh, Pittsburgh, PA, 2004.

[14] K. Nakayama et al., "Wikipedia mining - wikipedia as a corpus for knowledge extraction," *Proc. of Wikimedia International Conference (Wikimania)*, 2008.

[15] R. Nuray and F. Can, "Automatic ranking of information retrieval systems using data fusion," *Information Processing & Management*, vol.42 no. 3, pp.595-614, 2006.

[16] M. Ostberg, "Subtopic extraction using graph-based methods," M.S. thesis, Dept. Computer Science and Engineering, KTH Royal Institute of Technology, Sweden, 2013.

[17] L. Page et al., "The pagerank citation ranking: Bringing order to the web," Stanford InfoLab, Stanford University, California, Tech. Rep. TR-1999-66, 1999.

[18] D. Puspitaningrum et al., "An MDL-based frequent itemset hierarchical clustering technique to improve query search results of an individual search engine." *LNCS 9460*, pp.279-291, 2015.

[19] D. Puspitaningrum et al., "The analysis of rank fusion techniques to improve query relevance," *J. Telkomnika*, vol. 13 no. 4, pp.1495-1504, 2015.

[20] M. Shokouhi et al., "Query suggestion and data fusion in contextual disambiguation," *WWW 2015 Proceedings*, pp.971-980, 2015.

[21] G. Tsatsaronis et al., "An experimental study on unsupervised graph-based word sense disambiguation," in *Computational Linguistics and Intelligent Text Processing*, LNCS 6008, A. Gelbukh, Ed. Berlin: Springer, 2010, pp.184-198.

[22] N.C.Z. Yin and M. Shokouhi, "Query expansion using external evidence," in *ECIR 2009*, M. Boughanem et al., Eds. Heidelberg: Springer, 2009, pp.362-374.

Citation:
D. Puspitaningrum, G. Yulianti, I. S. W. B. Prasetya.
"Wiki-MetaSemantik: A Wikipedia-derived query expansion approach based on network properties" in 2017 5th International Conference on Cyber and IT Service Management (CITSM), Aug 2017, pp. 1-6. DOI: 10.1109/CITSM.2017.8089228. URL: http://ieeexplore.ieee.org/document/8089228/