Web Search Enhancement Using WordNet Query Expansion Technique*

Belal Al-Khateeb
Computer Science Dept.
College of CS and IT
University of Anbar
Ramadi, Iraq
belal@computer-college.org

Ali J. Hilal
Computer Science Dept.
College of CS and IT
University of Anbar
Ramadi, Iraq
ali_ja1991@yahoo.com

Sufyan Al-Janabi
[2]College of Science and Tech.,
UHD, Sulaimani, KRG-Iraq
[1]College of CS and IT
University of Anbar, Ramadi, Iraq
saljanabi@fulbrightmail.org

Abstract— **Search on the internet is a very important process for most people as many people use search engines to do research, obtain scientific documents, browse social networks (such as Facebook, twitter, etc.), email, etc. The user's ability to find useful information by using the right keywords in the query is considered as the main challenge in the search process. Sometimes, the user uses keywords that have more than one meaning where every meaning has more than one keyword to describe it. The proposed system aims to select the right meaning of the keywords and expanding the query to other keywords that have same meaning using WordNet. After the query expansion is done, Google search engine is used to do the search. The obtained results are promising, opening further research directions for enhancing the search process.**

Keywords- Information Retrieval, IR, Search Engine, SE, WordNet, Query Reformulation.

I. INTRODUCTION

Huge advances in computer technology have led to the start of what has been named as the information age. There is a very large amount of information on a variety of subjects available on networked media and internet. The challenge is retrieving useful information when it is required. This task should be done by an efficient and effective ways for organizing and indexing the data. The wide variety of forms in which information can be stored and transferred makes this task more challenging [1].

The concept of the information retrieval (IR) can be very wide. But on the whole, IR is the action of discovering stored information related to an information need from a group of stored information [2]. The IR system that is applied on the internet called search engine (SE). So, SE is a practical application of information retrieval system which helps people to find information on the World Wide Web [3].

The internet search engines and other modern information retrieval systems must apply the basic requirements; those requirements are: Firstly, there are no restrictions on the user when he/she wants to enter the query in the natural language, without the need to enter any operators. Secondly, the retrieved document should be ranked by degree of relevance to the query of the user. Thirdly, these systems must take into account the user feedback to the reformulation of the user query [4].

An IR system has three main processes which are: indexing of the documents, representation of the user's information need, and comparison between them. The indexing process include the following steps: take every document in the collection to be indexed, tokenize the text of the documents and each token called term, do linguistic preprocessing for each term, and index the terms produced from the process. This process is done in an off-line mode, so the IR system access to the representation of the document which is processed in advance. The IR system takes the query as input and the query attempt to describe the user's information need. The comparison process takes place between the document representation and the query [4] [5].

There are two popular measures used to evaluate the search engine: Precision and Recall. Precision is the fraction of the documents retrieved that are relevant to the information need and Recall is the fraction of the retrieved documents that are relevant to the query that are successfully retrieved [2]. The Precision specifies whether the documents retrieved are relevant and Recall specifies whether all the relevant documents are retrieved [6]. Precision describes the amount of the valid information in the search results. This reflects the helpfulness for users. Recall is used to describe the ratio of useful information in relative to all information in the search results that meet users' need [7]. Briefly, Recall retrieves all relevant documents (e.g. Legal) and Precision retrieves the most relevant documents (e.g. Web).

$$Precision = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \quad (1)$$

$$Recall = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{relevant\ documents\}|} \quad (2)$$

*This paper was presented at the Third International Scientific Conference of University of Human Development (April, 2016)

For example, if there are 10,000 documents in computer science, from those documents, there are 55 documents relevant to "Artificial intelligence" topic. Suppose that the search engine retrieves 40 documents (30 relevant and 10 irrelevant) after doing searches with keywords “Artificial intelligence”. Hence,

$$Precision = \frac{40 \cap 55}{40} = \frac{30}{40}$$

$$Recall = \frac{40 \cap 55}{55} = \frac{30}{55}$$

There is another important factor to evaluate search engine which is the speed of the search engine (the time spent to retrieve the results). The aim of IR algorithm is to try to obtain a higher precision rate and better recall rate in little time. However, this can be a difficult task. Therefore, it is a usual practice to try to reach a balance between them [8] [9].

The similarity of two documents can be measured using Jaccard similarity (JS). JS is a fraction of the size of the intersection to the size of the union of the documents sets [17].

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}. \quad (3)$$

Query expansion is the task of adding new terms to the original query to increase the number of retrieved documents which are relevant to user need. The significant problem of query expansion is the choice of the expansion keywords established on the original query, WordNet can be used to resolve this problem [10].

II. LITERATURE SURVEY

There are many kinds of literature that used different techniques for implementing query expansion in the search engine. Here are some earlier studies in the search engine and WordNet which are related to this work.

J. Nemrava [11] used WordNet Glosses to refine Google queries. He had chosen several techniques of how to organize returned web sites into proper synonym classes by using WordNet. He tested a group of 50 suitable nouns from some different areas. The central problem with the proposed system was the response time. It takes around 50 seconds with typical 20 Google queries for each one equivalent word class.

Ashish K. and Nitin C. [12] proposed a hybrid strategy for refining the search engine results through document clustering, Query Recommendation and genetic algorithm to supply the user with the top important pages to the search request. The proposed system starts with query recommendation, a genetic algorithm which is useful to resulting pages from query recommendation to guide maximum important pages to a user at a minimum time.

R. Hemayati et al [18] proposed technique assemblies the search results depending on the different meanings of the query. They proposed a novel three-step grouping algorithm that combines both categorization and clustering techniques. They also suggested an algorithm to combine like senses give back from WordNet. They achieved an accuracy of about 90%.

Abdelmgeid A. Aly [19] proposed an adaptive method using GA to adapt queries, based on relevance judgments. This algorithm was adapted for the three famous documents groups (CISI, NLP, and CACM). The algorithm showed the distinct effects of applying GA to increase the effectiveness of queries in IR systems. The aim was to recover most related documents with less number of irrelevant documents in information retrieval system using a genetic algorithm. The proposed GA approach gives better results than classical IR system when tried.

III. WORDNET

WordNet is a free online lexical reference system, whose design was inspired by some modern psycholinguistic theories of human lexical memory. Its synonym sets are well-organized depending on English nouns, verbs, and adjectives "the underlying lexical concepts". The synonym sets are connected by different relations. For making a dictionary which easy to deal with and proving the most benefits as much as possible. The initial idea was to use it in searching dictionaries conceptually, rather than merely alphabetically, so it was to be used as an on-line conventional dictionary with a closed conjunction. For those ideas, a group of psychologists and linguists at Princeton University in 1985 undertook the responsibility to improve a lexical database. The outcome was more ambitious of which were planned. This project has been named as "WordNet". Whereas it has been relying on the results of psycholinguistic research, WordNet can be said to be a dictionary based on psycholinguistic principles [13] [14] [15].

The WordNet divides the vocabulary into five classes: nouns, verbs, adjectives, adverbs, and function words. There are a lot of relations in WordNet such as [14][16]:

a) **Synonymy**: ***X*** will be synonymy of the ***Y***, when substitution ***X*** and ***Y*** can be done in any context and the truth value not change. For example, table and tabular array. Each set of synonyms is called *Synset* and given a unique *Synset ID* to identify it. Each *Synset* comes along with a brief description called a *gloss*. The glosses are usually one or two sentences long.

b) **Hyponymy**: ***X*** is a hyponym of ***Y***, when ***X*** is a kind of (type of) ***Y***. For example, Euphrates is the type of river. A hyponym is transitive and asymmetrical, a hyponym receives all the features of the most generic concept and adds at least one feature that makes it distinguished from its superordinate and from any other hyponyms of that superordinate. For example,

maple inherits the features of its superordinate, tree, but is distinguished from other trees by the hardness of its wood, the use of its sap for syrup, the shape of its leaves, etc.

c) **Meronymy**: ***X*** is a meronymy of ***Y***, if ***X*** is part of ***Y***. This relation is transitive (with qualifications) and asymmetrical. For example, an arm is a part of the body, a wheel is a part of the car.

d) **Antonym**: A concept which represents a word opposite is said to be Antonymy. This familiar concept turns out to be surprisingly difficult to define. But to simplify it, the logical formula of the term has been described. The antonym of a word ***x*** is sometimes not-x, but not always. Antonymy, which seems to be a simple symmetric relation, is actually completely complex; however, speakers of English have little difficulty recognizing antonyms when they see them.

With the comparison between WordNet and traditional dictionaries, many of the differences can be observed, e.g. the separation of the data into four databases associated with the categories of verbs, nouns, adjectives and adverbs. This organization system is justified by psycholinguistics research on the association of words to the syntactic categories by humans. Therefore, each database has been organized differently from the others. The names are organized in a hierarchy, the verbs by relations, the adjectives and the adverbs by N-dimension hyperspaces [13]. The main structure of the WordNet centers upon a word's semantics. The targeted looking-up for meaning-related words and concepts from multiple access points can be utilized. This property can be enabled using the digital format. The user can search for a keyword's Hyponym, Meronym, Antonym, Morphologically derived words, etc, by a browser with a pull-down menu. Unlike a traditional thesaurus such as Roget's, the arcs among WordNet's words and Synsets express a finite number of well-defined relations [16].

IV. PROPOSED SYSTEM MODEL

The proposed system has four basic parts: preprocessing stage, WordNet, Google Search engine and AI algorithms (see Figure 1). The preprocessing stage has three basic steps which are tokenization, dropping common terms, and lemmatization. Preprocessing stage aims to clean the query and obtain the basic terms in the query to do an efficient search. In tokenization step, the stream of text such as query is tokenized into a list of tokens or terms. The list of terms becomes an input to dropping common term step. In dropping common terms step, most frequently used words will be removed from the query (such as. Is, a, are, what, how and etc.). This step leads to speed up the system because the system removes some words that have a small value of significance and cannot help in retrieve process when matching the query with documents. In lemmatization step, derivational affixes of each term in the query will be eliminated in order to return the terms to the base or dictionary form.

WordNet is one of the basic parts of the proposed system model. The proposed system model uses WordNet to obtain the synonymy, polysemy, and glosses for each term. Through WordNet, the system can expand the query by adding new terms (synonymy of the terms) into the query. So, when the user enters the word "Rook" as a query, the WordNet displays the gloss of the term. Though this option we can select what we mean by "Rook" and use the synonymy of the word and gloss to expansion the query. However, the input to this step is the term and the output is set of synonymy and glosses to use it in the expansion the query.

Google search engine can be used to do the search of the original query, the query after expansion and the glosses of the keywords. After doing the search by Google, the results will be evaluated by computing the fitness. Fitness is a very important process used to rank the URLs in the result. The good results come from a proper selection of the fitness (see Figure 2). The proposed model parses every URL page and uses the meta-data of URL page. The system computes the fitness depending on:

a) Order of the URL page in Google results.

b) Percentage of similarity between the query and description of the URL page.

c) Percentage of similarity between the glosses of the query and description of the URL page.

d) Percentage of similarity between the query and title of the URL page.

The percentage of similarity can be computed by Jaccard score. Every one of these four factors has a different degree of importance. When using probability (0.1, 0.2, 0.3, 0.4) we obtain best results relevant to user query, because we obtain higher importance (0.4) to Similarity (Title of URL, Query), (0.3) to Similarity (Query, Description of URL), (0.2) to Similarity (Glosses of keywords, Description of URL) and lower importance (0.1) to order of URL in results. The value of fitness for each URL will be between 0 and 1.

AI algorithms can be used in several contexts. For example, genetic algorithm (GA) can be used to help the system in finding best results. GA represents the query and documents as a chromosome. The result of the navigate query in Google search engine will be population. GA generates the initial population for the query chromosome. Then the fitness will be computed for every URL by matching the query and document chromosome.

V. Results and Discussion

The proposed system is mainly designed for expanding the user's query by using other keywords that have same meaning using WordNet. 50 results from Google before and after applying the system are taken as shown in Table 1 and Figure 3. The obtained results clearly show that the precision before applying the system is low; this is due to ambiguous keywords that are found in the queries. The results in Table 1 show that the precision is very good when a search, after selecting the meaning of the keyword using WordNet, is done. Also, one can conclude that when AND operator between keywords in the query (after query expansion) is done then the precision in getting better, this is achieved only when the query has little keywords. In the same context, when we have a lot of keywords in the query (after query expansion) using OR operator in the query will give better precision than using AND operator.

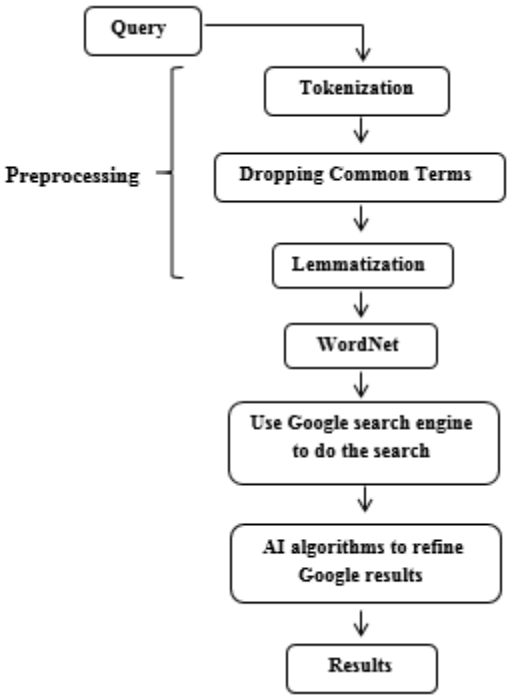

Figure 1: The Proposed System Model.

VI. Conclusion and Future Work

One can conclude that using query expansion via WordNet can successfully enhance the web search process. This success can indeed be increased when an AI algorithm (like GA) is introduced to the proposed system in order to help the system to find better results. This stage can be added after the proposed query expansion via WordNet. Further work will consider the effect using AI algorithms on query expansion and rank the results of the Google.

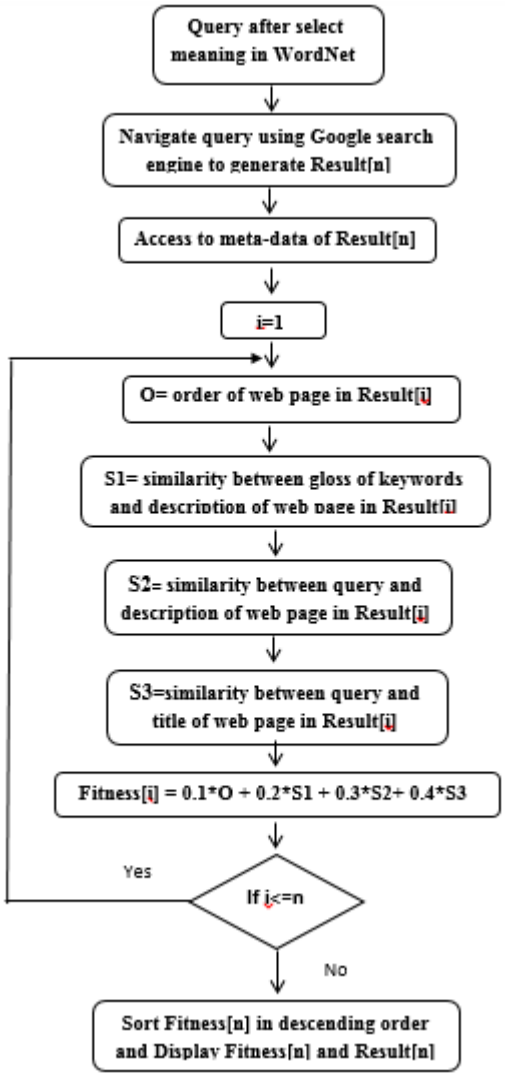

Figure 2: Results Evaluation.

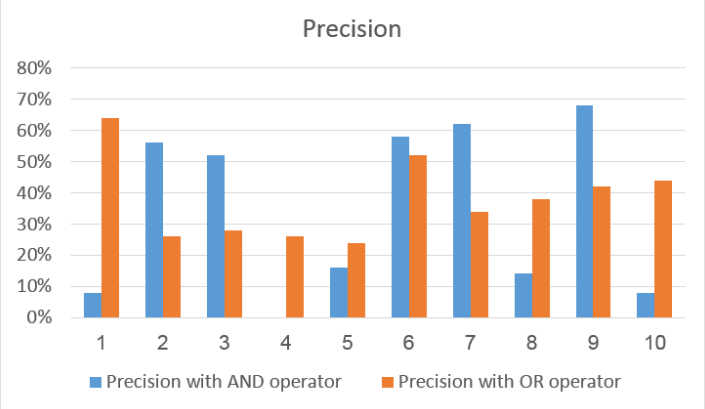

Figure 3: Precision with AND & OR operators.

REFERENCES

[1] M. Mitra and B. B. Chaudhuri, "Information Retrieval from Documents: A Survey", Information Retrieval, [Vol. 2, pp.141-163], 2000.

[2] S. Ceri, A. Bozzon, M. Brambilla, E. D. Valle, P. Fraternali, and S. Quarteroni, "Web Information Retrieval", Springer-Verlag Berlin Heidelberg, 2013.

[3] M. Levene, "An introduction to search engines and web navigation", John Wiley & Sons, 2011.

[4] D. Hiemstra, "Using language models for information retrieval", Taaluitgeverij Neslia Paniculata, 2001.

[5] D. Manning, P. Raghavan and H. Schütze. "Introduction to information retrieval", Cambridge: Cambridge university press, [Vol. 1], 2008.

[6] D. Minnie and S. Srinivasan, "Intelligent Search Engine algorithms on indexing and searching of text documents using text representation", IEEE International Conference in Recent Trends in Information Systems, [pp. 121-125], 2011.

[7] M. Gordon, Probabilistic and genetic algorithms for document retrieval, Communications of the ACM 31 (10) (1988) 1208–1218.

[8] J. Jiang, Z. Wang, C. Liu, Z. Tan, X. Chen and M. Li, "The Technology of Intelligent Information Retrieval Based on the Semantic Web", 2nd International Conference on Signal Processing Systems (ICSPS), IEEE, China, [Vol. 2, pp. 824-827], 2010.

[9] M. Kobayashi and K. Takeda, "Information retrieval on the web", ACM Computing Surveys (CSUR), [Vol. 32 No. 2, pp.144-173], 2000.

[10] H. Imran and A. Sharan, "Thesaurus and query expansion", international Journal of Computer science & Information Technology (IJCSIT), Vol 1, No 2, November 2009.

[11] Nemrava, Jan. "Using WordNet glosses to refine Google queries" In Proc. of the Dateso 2006 Workshop. VSB–Technical University of Ostrava, Dept. of Computer Science, pp. 85-94. 2006..

[12] Kushwaha, Ashish Kumar, and Nitin Chopde. "Hybrid Approach for Optimizing the Search Engine Result.", International Journal of Computer Science and Mobile Computing, Vol. 3, Issue. 4, pg.707 – 710 , April 2014.

[13] C. Fellbaum, "WordNet(s) In: Keith Brown", Elsevier, Encyclopedia of Language & Linguistics, [vol. 13, pp. 665-670], 2006.

[14] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database", International journal of lexicography, [Vol. 3, No. 4, pp. 235-244], 1990.

[15] N. Polettini, "The vector space model in information retrieval-term weighting problem." Entropy, [pp. 1-9], 2004.

[16] Z. Elberrichi, A. Rahmoun, and M. A. Bentaalah, "Using WordNet for Text Categorization", International Arab Journal of Information Technology, [Vol. 5, No. 1, pp. 16-24], 2008.

[17] G. Miller, "WordNet: A Lexical Database for English", Communications of the ACM,.[Vol 38, No 11,pp. 39-41], 1995

[18] Hemayati, Reza, Weiyi Meng, and Clement Yu. "Semantic-based grouping of search engine results using WordNet." Advances in Data and Web Management. Springer Berlin Heidelberg, 2007. 678-686.

[19] Abdelmgeid A.Aly, "Applying genetic algorithm in query improvement problem", International journal "Information Technologies and Knowledge" Vol 1, 309 - 316, 2007.

Table 1: List of Queries Before and After Applying the System.
(RL is the Number of Relevant Pages, RT is the Number of Retrieved pages, and P is the Precision)

						Using AND Operator			**Using OR Operator**		
No.	**Original Query**	**RL**	**RT**	**P**	**Select meaning using WordNet**	**RL**	**RT**	**P**	**RL**	**RT**	**P**
1	Hear	2	50	4%	**learn, hear, get word, get wind, pick up, find out, get a line, discover, see** -- (get to know or become aware of, usually accidentally; "I learned that she has two grown-up children"; "I see that you have been promoted")	4	50	8%	32	50	64%
2	Line	4	50	8%	**course, line** -- (a connected series of events or actions or developments; "the government took a firm course"; "historians can only point out those lines for which evidence is available")	28	50	56%	13	50	26%
3	Hate sport	6	50	12%	**hate, detest** -- (dislike intensely; feel antipathy or aversion towards; "I hate Mexican food"; "She detests politicians") **sport, feature, boast** -- (wear or display in an ostentatious or proud manner; "she was sporting a new hat")	26	50	52%	14	50	28%

4	Hood	9	50	18%	**hood, bonnet, cowl, cowling** -- (protective covering consisting of a metal part that covers the engine; "there are powerful engines under the hoods of new cars"; "the mechanic removed the cowling in order to repair the plane's engine")	21	50	42%	13	50	26%
5	Foot	3	50	6%	**foundation, base, fundament, foot, groundwork, substructure, understructure** -- (lowest support of a structure; "it was built on a base of solid rock"; "he stood at the foot of the tower")	8	50	16%	24	50	24
6	Mouse	1	50	2%	**shiner, black eye, mouse** -- (a swollen bruise caused by a blow to the eye)	29	50	58%	21	50	52%
7	Butt	5	50	10%	**target, butt** -- (sports equipment consisting of an object set up for a marksman or archer to aim at)	31	50	62%	17	50	34%
8	Mole	8	50	16%	**breakwater, groin, mole, bulwark, seawall, jetty** -- (a protective structure of stone or concrete; extends from shore into the water to prevent a beach from washing away)	7	50	14%	19	50	38%
9	Center Java	20	50	40%	**center, centre, middle, heart, eye** -- (an area that is approximately central within some larger region; "it is in the center of town"; "they ran forward into the heart of the struggle"; "they were in the eye of the storm") **Java** -- (an island in Indonesia to the south of Borneo; one of the world's most densely populated regions)	34	50	68%	21	50	42%
10	Good Note	3	50	6%	**adept, expert, good, practiced, proficient, skillful, skilful** -- (having or showing knowledge and skill and aptitude; "adept in handicrafts"; "an adept juggler"; "an expert job"; "a good mechanic"; "a practiced marksman"; "a proficient engineer"; "a lesser-known but no less skillful composer"; "the effect was achieved by skillful retouching") **note, musical note, tone** -- (a notation representing the pitch and duration of a musical sound; "the singer held the note too long")	4	50	8%	22	50	44%