Enhanced Google-Based Semantic Category Search

Mohammed Adeeb[1], Ahmed Sleman[1], Sumaya Abdullah[1] *and* **Belal Al-Khateeb[1]**

1 Department of Computer Science, College of Computer, Al-Anbar University, Iraq

***Abstract*— Recently search services have been developed rapidly especially when the social internet appeared. It can help web users easily find their documents. So that it is very difficult to find a best search method. This paper aims to enhance the quality of the search engines results and this can be done by adding a second level category search that is able to search for the keyword and its synonyms, which enables the search engines to get more users queries related results. The proposed method showed promising results that will open further research directions.**

***Index Terms*— Search Engine, Google, Information Retrieval, Keyword, Category.**

I. INTRODUCTION

Recently, the user's satisfaction of search engine results is decreased because search engines become more critical for finding information over the World Wide Web where web content growing fast. Also, the search engines return a huge number of web pages, and then the user may take long time to look at all of these pages to find his needed information, the difficulty of having a right query, the difficulty of knowing which results are similar and so on, due to the recent improvements of search engines and the rapid growth of the web [1]. Currently, techniques for content description and query processing in Information Retrieval (IR) are based on keywords, and therefore provide limited capabilities to capture the conceptualizations associated with user needs and contents. Aiming to solve the limitations of keyword-based models, the idea of conceptual search, understood as searching by meanings rather than literal strings, has been the focus of a wide body of research in the IR field. [2]. A good search engine should contain as many relevant, high-quality pages and as few irrelevant, low quality pages as possible. It is hard to build a comprehensive and relevant collection for a search engine; this is due to web's large size and diversity of content.

Spiders can be used by search engines usually use spiders to retrieve pages from the web by recursively following URL links in pages using standard HTTP protocols. These spiders (also referred to as Web robots, crawlers, worms, or wanderers) use different algorithms to control their search, the following methods have been used to locate web pages that are relevant to a particular domain; The spiders can be restricted to staying in particular web domains, because many web domains have specialized contents. While some spiders are restricted to collecting only pages at most a fixed number of links away from the starting URLs or starting domains. Assuming that nearer pages have higher chances of being relevant, this method prevents spiders from going too "far away" from the starting domains. Finally more sophisticated spiders use more advanced graph search algorithms that analyze Web pages and hyperlinks to decide what documents should be downloaded.

In most cases the resulting collection is still noisy and needs further processing. Filtering programs are needed to eliminate irrelevant and low-quality pages from the collection to be used in a search engine. There are four different filtering techniques that can be used to eliminate such a noise from the obtained search results; Domain experts manually determine the relevance of each Web page (e.g., Yahoo). In the simplest automatic procedure, the relevance of a Web page can be determined by the occurrences of particular keywords. Web pages are considered relevant if they contain the specified keyword, and are considered irrelevant otherwise. TFIDF (term frequency inverse document frequency) is calculated based on a lexicon created by domain experts. Web pages are then compared with a set of relevant documents, and those with a similarity score above a certain threshold are considered relevant.

Text classification techniques such as the Naïve Bayesian classifier also have been applied to Web page filtering [3]. It is worth to mention that some search engines do not perform filtering; they assume that most pages found in the starting domains (or at a specified depth) are relevant [4]. This paper aims to improve the efficiency of specific search engines in locating the URLs that point to relevant Web pages. This can be done by using a second level category search and finding the occurrences of particular keywords and its synonyms in the search results. Relevant web pages are considered if they contain the specified keyword or one of its synonyms, otherwise it will be considered as irrelevant. It is worth to mention that the work in this paper represents an extension to the work done by Al-Khateeb et.al. [5].

However, the most common use is the one as an improved form of search on the Web, where meaning and structure are extracted from both the user's Web search queries and different forms of Web content, and exploited during the Web search process. Such semantic search is often achieved by using Semantic Web technology for interpreting Web search queries and resources relative to one or more underlying ontologies, describing some background domain knowledge, in particular, by connecting the Web resources to semantic annotations, or

by extracting semantic knowledge from Web resources. [6]. A branch of information retrieval research focuses on techniques that improve the accuracy of search results. One such technique is query difficulty prediction. Query difficulty prediction is the task of determining the effectiveness of search without any further information about the query from the user. It is difficult to predict query difficulty and this expected because it involves natural language so it is not always easy to know what the user want. A query can be difficult because a user does not provide enough information, or because the query itself has a complex meaning that a token-based search system fails to understand [7]. Query expansion technique is used to improve the correctness of a search engine. This can be done by attaching additional concepts to the search query of the user. These attached concepts could be user specific information or the expansion of the query with synonyms, hypernyms or hyponyms. [8]. Another method that can be used in the web pages retrieval is the keyword-spice method; this method considers those web pages that contain the user's input query keyword only and not all the web pages. [9,10]. The semantic modification of user queries is a well-known technique from information retrieval. In the area of semantic search it often exploits information from ontologies. It plays a central role in many semantic search engines. Different techniques have been developed to increase both, recall and precision of a query [1].

The query language of a standard search engine is simply a list of keywords. In some search engines, each keyword can optionally be prepended by a plus sign ("+"). Keywords with a plus sign must appear in a satisfying document, whereas keywords without a plus sign may or may not appear in a satisfying document (but the appearance of such keywords is desirable [11]. The search results of the Google Search Engine will be different according to the arrangement of keywords in the search query. As the novice web users are not familiar with the construction of effective keywords for their search queries, Guided Google provides a function that will automatically calculate the permutation and make different combinations of the keywords used. In Google search, the words in quotes mean that they have to occur in that particular order, in the search results. So that if the search query is placed in quotes, the result of the combinations will also be reflected in quotes [12].

II. BACKGROUND

Currently, techniques for content description and query processing in Information Retrieval (IR) are based on keywords, and therefore provide limited capabilities to capture the conceptualizations associated with user needs and contents. Aiming to solve the limitations of keyword-based models, the idea of conceptual search, understood as searching by meanings rather than literal strings, has been the focus of a wide body of research in the IR field. In [13] the authors reported their research on utilizing semantic model to improve the searching function within Spatial Web Portals (SWPs). Based on SWEET, they built the domain ontology and implemented a semantic inference service. Multiple data resources are bridged to provide cross catalog searches, and to support spatial search in an intelligent manner. In [`6] Bettina Fazzinga and Thomas Lukasiewicz give a brief overview of existing such approaches, including own ones, and sketch some possible future directions of research. Some of the most pressing research issues are maybe (i) how to automatically translate natural language queries into formal ontological queries, and (ii) how to automatically add semantic annotations to Web content, or alternatively how to automatically extract knowledge from Web content. Another central research issue in semantic search on the Web is (iii) how to create and maintain the underlying ontologies. In [14] the same authors and others present a novel approach to Semantic Web search, which is based on ontological conjunctive queries, and which combines standard Web search with ontological background knowledge. Showing how standard Web search engines can be used as the main inference motor for processing ontology-based semantic search queries on the Web. Miriam Fernández et. al [2] investigate the definition of an ontology-based IR model, oriented to the exploitation of domain Knowledge Bases to support semantic search capabilities in large document repositories, stressing. In [15] Lukasiewicz et. al argued that such rankings can be based on ontological background knowledge and on user preferences. Another aspect that has become increasingly important in recent times is that of uncertainty management, since uncertainty can arise due to many uncontrollable factors. To combine these two aspects, they proposed extensions of the Datalog+/– family of ontology languages that both allow for the management of partially ordered preferences of groups of users as well as uncertainty, which is represented via a probabilistic model. In [16] the same authors describe how to combine ontological knowledge with CP-nets to represent preferences in a qualitative way and enriched with domain knowledge. Specifically, they focus on conjunctive query (CQ) answering under CP-net-based preferences. They have defined k-rank answers to CQs based on the user's preferences encoded in an ontological CP-net and they have provided an algorithm for k-rank answering CQs. In [17] S. Anuradha et. al have presented an automatic annotation approach that first aligns the data units on a result page into different groups such that the data in the same group have the same semantic. Then, for each group they have annotated it from different aspects and aggregate the different annotations to predict a final annotation label for it. An annotation wrapper for the search site is automatically constructed and can be used to annotate new result pages from the same web database. In [18] the authors tackled the problem of query

answering in Datalog+/– ontologies subject to the querying user's preferences and a collection of subjective reports (i.e., scores for a list of features) of other users, who have their own preferences as well. All these pieces of information are combined to rank the query results. They first focus on the problem of ranking atoms in a database by leveraging reports and customizing their content according to the user's preferences. Then, they extend this approach to deal with ontological query answering using provenance information. The use of Datalog+/– for information integration based on probabilistic data exchange. More specifically, studying the previously introduced probabilistic data exchange problem consisting of a probabilistic database as a source, source-to-target mappings in Datalog+/– and a target Datalog+/– ontology. A complexity analysis was provided for deciding the existence of (deterministic and probabilistic (universal)) solutions in the context of data exchange. In particular, tractability is preserved for simple probabilistic representations, such as tuple-independent ones was showed [19].

III. THE PROPOSED SEARCH METHOD

In this work we applied a second level search in order to find better results and more relative web pages based on user queries. This was done by using additional keywords, or vocabulary that refers to the field or category which the user query belongs to. This method is implemented in four different ways, those are a keyword that is point to single category, a keyword that have more than one category, the use of vocabulary with synonymous or use a description for the category. The process of our search retrieves web pages using category (keyword) search and compare it with the results of Google with/without using category keyword. The following steps are used to get such results:

1- Get raw search results: by taking the search query and the keyword from the user then downloading:
   a. Google search page without the keyword for 100 results.
   b. Google search page with the keyword for 10 results.
2- Parsing these two search results pages: this step decomposes search results into (Title, URL, Description and the Repetition of the keyword in the title and description)
3- Processing Google with the keyword (Gwith):
   a. Set max =0 , found =0 , proc =0
   b. for each result in Gwith do steps c through e:
   c. get the synonyms list by searching for keyword in the synonyms database.
   d. If this result contains the keyword or one of its synonyms then
   - increase proc
   else skip this result.
   e. Search for this result in Gwithout, if found and the no. of this result is larger than max then
   - max=result no.
   - increase found.
   f. Output: "Gwithout needed: "max" results to fulfill: "found" out of: "proc" from Gwtih".
4- Processing Google without (Gwithout) the keyword:
   a. Set max=0, bound =0
   b. While (bound < 100) or (max ==10) do step c
   - If the current search result (from Gwithout) contains the keyword or one of its synonyms then increase max.
   - Increase bound
   c. Output: "Our search found: "max" results containing the keyword inbound of: "bound" from Gwithout".

IV. RESULTS

The above algorithm is tested in different cases (ten different searches for each case) and the percentages of the results are calculated in order to measure the efficiency of the proposed search. Sections *a* through *e* show the obtained results.

A. Using Keyword with Category

In this case we used a keyword point to a single category as a second level of search. The results are shown in table I.Table I: Single Word Category Search.

Keyword	Category	Gwith vs Gwithout	Percent	Oursearch vs Gwithout	Percent
Galaxy Note 2	Mobile	10 vs 99	10.10 %	10 vs 36	27.77 %
router	Network	10 vs 99	10.10 %	10 vs 18	55.55 %
software	computer	10 vs 99	10.10 %	10 vs 66	15.15 %
Game	Kids	10 vs 99	10.10 %	10 vs 64	15.62 %
hepatitis	Viral	10 vs 99	10.10 %	10 vs 23	43.47 %
hemothorax	Trauma	10 vs 17	58.82 %	10 vs 10	100 %
clotting	Bleeding	10 vs 48	20.83 %	10 vs 33	30.30 %
ford	Car	10 vs 99	10.10 %	10 vs 45	22.22 %
search	Engine	10 vs 75	13.33	10 vs 23	43.47 %
engines	Car	10 vs 99	10.10 %	10 vs 62	16.12 %

Table I showed that ten out of ten results in our search is better than Google with the second level search, which is considered as a clear success for our proposed search algorithm.

B. Using Keyword That Have Two Categories

Table II shows the results of using single word keywords that belong to two categories.

Table II: Single Word with Two Categories Search.

Keyword	Category	Gwith vs Gwithout	Percent	Oursearch vs Gwithout	Percent
apple	Company	10 vs 99	10.10 %	9 vs 100	9 %
apple	Fruit	10 vs 99	10.10 %	1 vs 100	1 %
Sony	computers	10 vs 99	10.10 %	1 vs 100	1 %
sony	Tv	10 vs 99	10.10 %	10 vs 22	45.45 %
panda	Bear	10 vs 99	10.10 %	8 vs 100	8 %
panda	antivirus	10 vs 99	10.10 %	4 vs 100	4 %
photography	Photo	10 vs 99	10.10 %	10 vs 10	100 %
photography	Art	10 vs 99	10.10 %	10 vs 25	40 %
computer	science	10 vs 99	10.10 %	10 vs 51	19.6 %
computer	Pc	10 vs 99	10.10 %	10 vs 70	14.28 %

It is clear that our proposed search beats Google with the second level search in five out of ten results, which is considered as a fair success for our proposed search algorithm.

C. Using Keyword with Synonyms of Category

In this case we used a keyword point to a single category with synonyms as a second level of search. The results are shown in table III.

We see that six out of ten results in our search are better than Google with the second level search, which is considered as an acceptable success for our proposed search algorithm.

D. Using Sentence with Category

In this case we use a sentence keyword point to a single category in the second level of search. Table IV shows the obtained results.

Table IV showed that seven out of ten results in our search is better than Google with the second level search, which is considered as a clear success for our proposed search algorithm.

Table III: Single Word Category with Synonyms Search.

Keyword	Category	Gwith vs Gwithout	Percent	Oursearch vs Gwithout	Percent
bmw	Car	10 vs 99	10.10 %	10 vs 23	43.47 %
bmw	Motor	10 vs 99	10.10 %	10 vs 18	55.55 %
melanoma	cancer	10 vs 99	10.10 %	10 vs 10	100 %
melanoma	malignant	10 vs 32	31.25 %	10 vs 77	12.98 %
hydrocortisol	cortisol	10 vs 36	27.77 %	10 vs 10	100 %
hydrocortisol	steroid	10 vs 36	27.77 %	7 vs 100	7 %
contusion	injury	10 vs 64	15.62 %	10 vs 34	29.41 %
contusion	trauma	10 vs 64	15.62 %	10 vs 34	29.41 %
nodule	lump	10 vs 12	83.33 %	10 vs 64	15.62 %
nodule	mass	10 vs 73	13.69 %	4 vs 100	4 %

Table IV: Single Sentence Category Search.

Keyword	Category	Gwith vs Gwithout	percent	Oursearch vs Gwithout	percent
What do vegans eat?	food	10 vs 11	90.90 %	10 vs 14	71.42 %
What is the capital of iraq?	city	10 vs 16	62.5 %	10 vs 10	100 %
Eye color: the family genes?	Genetics	10 vs 99	10.10 %	10 vs 13	76.92 %
How does Google rank your page?	Search engine	10 vs 33	30.30 %	10 vs 66	15.15 %
How to plan your site structure with keyword research	search	10 vs 10	100 %	10 vs 19	52.63 %
microsoft internet software	software	10 vs 11	90.90 %	10 vs 10	100 %
repair computer sound	computer	10 vs 15	66.66 %	10 vs14	71.42 %
human resources employment	jobs	10 vs 52	19.23 %	10 vs 10	100 %
panasonic home electronics	electronics	10 vs 21	47.61 %	10 vs 14	71.42 %
What is the best online game for iPod Touch?	game	10 vs 15	66.66 %	10 vs 10	100 %

E. Using Keyword with Category after adding "s" to the Keyword or Category

In this case we use different queries from above tables but we added "s" either with query or with a keyword that point to a single category in the second

level of search. The obtained results are shown in table V.

Table V showed that seven out of ten results in our search is better than Google with the second level search.

Table V: Single Word Category Search with "s"

Keyword	Category	Gwith vs Gwithout	Percent	Oursearch vs Gwithout	Percent
Galaxy Note 2	mobiles	10 vs 99	10.10 %	10 vs 69	14.49 %
Bmw	cars	10 vs 99	10.10 %	10 vs 36	27.77 %
router	networks	10 vs 99	10.10 %	10 vs 17	58.82 %
software	computers	10 vs 99	10.10 %	3 vs 100	3 %
ford	cars	10 vs 99	10.10 %	10 vs 53	18.86 %
search	engines	10 vs 77	12.98 %	6 vs 100	6 %
repair computer sound	computers	10 vs 33	30.30 %	10 vs 41	24.39 %
computers	science	10 vs 99	10.10 %	10 vs 92	10.86 %
panasonic home electronics	electronic	10 vs 17	58.82 %	10 vs 14	71.42 %
photography	photos	10 vs 99	10.10 %	10 vs 10	100 %

V. CONCLUSIONS AND RECOMMENDATIONS

This paper showed the enhancement of search engines by adding a second level category search. The method is implemented and tested in various cases and the results were promising. The results indicated that adding a second level category search will give better related results as our method outperformed Google in ten out of ten single word category search as shown in table I. Also our method was able to get better results in seven out of ten sentence category search and single word category search with "s" is added to either the keyword or to the category compared to Google as shown in table IV and V. The results in table III showed that our method is slightly better than Google in single word category with synonyms search. While both our method and Google had an equal performance in single word with two categories search as shown in tables II. It is worth to mention that our method used English text only results and ignored the results that may came in other languages that Google can fetch. Also our method ignored the video or images results.

So considering many popular languages and video and images results as a future work can enhance the obtained results. Also considering many samples as a future work can give a wider idea about the efficiency of the proposed method.

REFERENCES

[1] M. Christoph , "A survey and classification of semantic search approaches", Int. J. Metadata, Semantics and Ontology, Vol. 2, No. 1, 2007 23.

[2] F. Miriam, C. Iván, L.Vanesa, V. David, C. Pablo, and M. Enrico, " Semantically enhanced Information Retrieval: an ontology-based approach", Journal of Web Semantics: PREPRINT SERVER, Vol 9, No 4 (2011).

[3] M. Chau,, H. Chen, "A machine learning approach to web page filtering using content and structure analysis",Decision Support Systems 44 (2008) 482–494.

[4] M. Chau, Z. Huang, J. Qin, Y. Zhou, H. Chen, "Building a scientific knowledge web portal: the nanoport experience", Decision Support Systems 42 (2) (2006) 1216–1238.

[5] Al-Khateeb B., Abdullah S. and Adeeb M., A Proposed Google-Based Category Search, International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 3, Issue 9, September 2013.

[6] F. Bettina, L. Thomas, "Semantic search on the Web", Semantic Web 1 (2010) 89–96 89, DOI 10.3233/SW-2010-0023, IOS Press.

[7] Steven Garcia B.App.Sci. (Hons.), "Search Engine Optimisation Using Past Queries", School of Computer Science and Information Technology, Melbourne, Victoria, Australia. March 30, 2007.

[8] M. Robert, "Text-Mining for Semi-Automatic Thesaurus Enhancement", Diploma Thesis, August 2009.

[9] O. Satoshi, K. Takashi, I. Toru, "Keyword Spices: A New Method for Building Domain-Specific Web Search Engines", Venue: In Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01), Citations: 10 - 3 self, 2001.

[10] M. Andrew, N. Kamal, R. Jason, S. Kristie, "A Machine Learning Approach to Building Domain-Specific Search Engines", Venue: In Proceedings of the 16th International Joint Conference on Artificial Intelligence, Citations: 68 – 3 self, 1999.

[11] C. Sara , M. Jonathan , K. Yaron, S. Yehoshua, "XSEarch: A Semantic Search Engine for XML", Proceedings of the 29th VLDB Conference, Berlin, Germany, 2003.

[12] H. Choon Ding and B. Rajkumar ," Guided Google: A Meta Search Engine and its Implementation using the Google Distributed Web Services", arXiv.org > cs > arXiv:cs/0302018, Submitted on 13 Feb 2003.

[13] L. Wenwen, Y. Chaowei, R. Rob, "A Semantic Enhanced Search for Spatial Web Portals", AAAI Spring Symposium: Semantic Scientific …, 2008.

[14] F. Bettina Fazzinga, G. Giorgio, G. Georg, L. Thomas, " Semantic Web Search Based on Ontological Conjunctive Queries", Foundations of Information and Knowledge Systems, Volume 5956, 2010, pp 153-172

[15] L. Thomas, M. Maria Vanina, S. Gerardo I, and T. Oana Marciuska, "Ontology−Based Query Answering with Group Preferences", No. RR−14−02. DCS. May, 2014.

[16] N. Tommaso Di, L. Thomas, M. Maria Vanina, S. Gerardo I, and T. Oana Marciuska, "Computing k-Rank Answers with Ontological CP-Nets", Proceedings of the 22nd Italian Symposium on Advanced Database Systems, SEBD 2014, Sorrento Coast, Italy, June 16−18, 2014.

[17] S. Anuradha, G. Bhavana, Ch. D. V. Sudheer, G. Suryanarayana, G. Govind, " Enhanced Search Results with Semantic Annotation Approach", International Journal of Scientific & Engineering Research, Volume 5, Issue 6, June-2014.

[18] L. Thomas, M. Maria Vanina, M. Cristian, P. Livia and S. Gerardo, "Answering Ontological Ranking Queries Based on Subjective Reports", Proceedings of the 1st Workshop on Logics for Reasoning about Preferences, Uncertainty, and Vagueness, PRUV 2014, Vienna, Austria, July 23−24, 2014. Vol. 1205 of CEUR Workshop Proceedings. Pages 127−140. CEUR−WS.org. 2014.

[19] L. Thomas, M. Maria Vanina, P. Livia and S. Gerardo, " Information Integration with Provenance on the Semantic Web via Probabilistic Datalog+/", No. RR−15−01. DCS. 2015.