Volume 3, Issue 9, September 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

A Proposed Google-Based Category Search

Belal Al-Khateeb* *Department of Computer Science College of Computer Al-Anbar University, Iraq*

Sumaya Abdullah *Department of Computer Science College of Computer Al-Anbar University, Iraq*

Mohammed Adeeb *Department of Computer Science College of Computer Al-Anbar University, Iraq*

Abstract*— Search services have been developed rapidly in social internet. 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 search engines results by adding a second level category search, which enables the search engines to get more users queries related results. The proposed method showed promising results that will open further research directions.

Keywords*— Search Engine, Google, Information Retrieval, Keyword, Category.

I. INTRODUCTION

In the period of internet, search engine is the popular and necessary tool for the web users. However, people are often not satisfied when using them because of the difficulty of choosing the right results from the huge list of results, the difficulty of having a right query, the difficulty of knowing which results are similar and so on [1]. Information Retrieval (IR) is generally defined as the gathering of items that satisfy the user's need of information. In general this means that IR does not claim to answer specific questions but to retrieve documents which include the answer for this question. By the dispersion of the internet and the thereby enormous increase of published and available documents, efficient and automatic methods for IR become necessary [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 in the search results. Relevant web pages are considered if they contain the specified keyword, otherwise it will be considered as irrelevant.

II. BACKGROUND

Building a domain-specific web search engine requires making indices to domain documents and this can be done by running web-crawling spiders that collect only relevant pages [5]. 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 [6]. 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. [2]. 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. [7,8]. 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 [9]. 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 [10].

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 and d:
   c. If this result contains the keyword then
      - increase proc

      else skip this result.
   d. Search for this result in Gwithout, if found and the no. of this result is larger than max then
      - max=result no.
      - increase found.
   e. 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 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 80*	*12.5 %*	*10 vs 69*	*14.49 %*
Router	*network*	*10 vs 48*	*20.83 %*	*10 vs 18*	*55.55 %*
Software	*computer*	*10 vs 66*	*15.15 %*	*10 vs 33*	*30.30 %*
Game	*Kids*	*10 vs 61*	*16.39 %*	*10 vs 57*	*17.54 %*

Hepatitis	*Viral*	*10 vs 69*	*14.49 %*	*10 vs 46*	*21.73 %*
Hemothorax	*Trauma*	*10 vs 37*	*27.02 %*	*10 vs 32*	*31.25 %*
Clotting	*bleeding*	*10 vs 96*	*10.41 %*	*10 vs 47*	*21.27 %*
Ford	*Car*	*10 vs 55*	*18.18 %*	*10 vs 26*	*38.46 %*
Search	*Engine*	*10 vs 66*	*15.15 %*	*10 vs 40*	*25 %*
Engines	*Car*	*10 vs 69*	*14.49 %*	*10 vs 85*	*11.76 %*

Table I showed that nine 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 %*	*10 vs 58*	*17.24 %*
Apple	*Fruit*	*10 vs 99*	*10.10 %*	*2 vs 100*	*2 %*
Sony	*computers*	*10 vs 55*	*18.18 %*	*10 vs 32*	*31.25 %*
Sony	*Tv*	*10 vs 18*	*55.55 %*	*10 vs 25*	*40%*
Panda	*bear*	*10 vs 99*	*10.10 %*	*3 vs 100*	*3 %*
Panda	*antivirus*	*10 vs 79*	*12.65 %*	*10 vs 69*	*14.49 %*
photography	*Photo*	*10 vs 91*	*10.98 %*	*10 vs 10*	*100 %*
photography	*Art*	*10 vs 22*	*45.45 %*	*10 vs 35*	*28.57 %*
Computer	*Science*	*10 vs 66*	*15.15 %*	*10 vs 44*	*22.72 %*
Computer	*Pc*	*10 vs 27*	*37.03 %*	*10 vs 43*	*23.25 %*

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 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

.

Table III
Single Word Category with Synonyms Search

Keyword	***Category***	***Gwith vs Gwithout***	***Percent***	***Oursearch vs Gwithout***	***percent***
Bmw	*Car*	*10 vs 31*	*32.25 %*	*10 vs 27*	*37.03 %*
Bmw	*Motor*	*10 vs 23*	*43.47 %*	*10 vs 29*	*34.48 %*
Melanoma	*Cancer*	*10 vs 80*	*12.5 %*	*10 vs 14*	*71.42 %*
Melanoma	*malignant*	*10 vs 75*	*13.33 %*	*10 vs 90*	*11.11 %*
hydrocortisol	*cortisol*	*10 vs 19*	*52.63 %*	*10 vs 10*	*100 %*
hydrocortisol	*Steroid*	*10 vs 95*	*10.52 %*	*6 vs 100*	*6 %*
Contusion	*Injury*	*10 vs 78*	*12.82 %*	*10 vs 35*	*28.57 %*
Contusion	*Trauma*	*10 vs 78*	*12.82 %*	*8 vs 100*	*8 %*
Nodule	*Lump*	*10 vs 91*	*10.98 %*	*10 vs 81*	*12.34 %*
Nodule	*Mass*	*10 vs 72*	*13.88 %*	*10 vs 81*	*12.34*

We see that five out of ten results in our search are better than Google with the second level search, which is considered as a 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
Single Sentence Category Search

Keyword	Category	Gwith vs Gwithout	percent	Oursearch vs Gwithout	percent
What do vegans eat?	Food	10 vs 25	40 %	10 vs 19	52.63 %
What is the capital of iraq?	city	10 vs 30	33.33 %	10 vs 15	66.66 %
Eye color: the family genes?	Genetics	10 vs 46	21.73 %	10 vs 24	41.66 %
How does Google rank your page?	Search engine	10 vs 95	10.52 %	10 vs 61	16.39 %
How to plan your site structure with keyword research	search	10 vs 36	27.77 %	10 vs 10	100 %
microsoft internet software	software	10 vs 99	10.10 %	10 vs 13	76.92 %
repair computer sound	computer	10 vs 99	10.10 %	10 vs10	100 %
human resources employment	jobs	10 vs 51	19.6 %	10 vs 23	43.47 %
panasonic home electronics	electronics	10 vs 36	27.77 %	10 vs 11	90.90 %
What is the best online game for iPod Touch?	Game	10 vs 34	29.41 %	10 vs 10	100 %

Table IV 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.

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: Single Word Category Search with “s”

Keyword	Category	Gwith vs Gwithout	Percent	Oursearch vs Gwithout	Percent
Galaxy Note 2	Mobiles	10 vs 23	43.47 %	2 vs 100	2 %
Bmw	Cars	10 vs 31	32.25 %	10 vs 42	23.8 %
Router	Networks	10 vs 33	30.30 %	10 vs 77	12.98 %
Software	computers	10 vs 93	10.75 %	3 vs 100	3 %
Ford	Cars	10 vs 69	14.49 %	10 vs 28	35.71 %
Search	Engines	10 vs 93	10.75 %	9 vs 100	9 %
repair computer sound	computers	10 vs 99	10.10 %	10 vs 52	19.23 %
Computers	Science	10 vs 46	21.73 %	6 vs 100	6 %
panasonic home electronics	Electronic	10 vs 11	90.90 %	10 vs 11	90.90 %
photography	Photos	10 vs 92	10.86 %	10 vs 88	11.36 %

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

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 was able to get better results in nine out of ten single word category search compared to Google as shown in table I. Also our method outperformed Google in ten out of ten sentence category search as shown in table IV. While both our method and Google had an equal performance in single word with two categories search and single word category with synonyms search as shown in tables II and III. The results in table V showed that Google is slightly better than our method in single word category search with "s" is added to either the keyword or to the category. 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] M. Robert, "Text-Mining for Semi-Automatic Thesaurus Enhancement", Diploma Thesis, August 2009.

[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] K. Seema, K. Narender, "Creating Topic Hierarchy With Clustering in Domain Specific Search Engine", IJESAT, Mar-Apr 2012.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

