APPROXIMATING USER'S INTENTION FOR SEARCH ENGINE QUERIES

Aya Awad[1], Maged El-Sayed[2] and Y. El-Sonbaty[3]

[1]*Business Information Systems Department, College of Management and Technology, Arab Academy for Science, Technology & Maritime Transport, Alexandria, Egypt*

[2]*Department of Information Systems and Computers, Faculty of Commerce, Alexandria University, Alexandria, Egypt*

[3]*College of Computing & I.T., Arab Academy for Science, Technology & Maritime Transport, Alexandria, Egypt*

Keywords: Information Retrieval, Semantic Search, Search Engine, User Intention.

Abstract: Documents on the internet are not organized in a way that eases search and retrieval by users using search engines. The user of the search engine is typically overwhelmed by the size of the returned result and does not normally look beyond the first few pages of result. Knowing that the majority of search engines are term-based, this information retrieval problem is caused by two issues: (1) query articulation issue; where the user is not capable of expressing his information need well, and (2) semantic gap issue where the search engine may not be able to retrieve semantically relevant documents. In this paper we introduce a solution that addresses these issues through semantic enrichment and query reformulation. Our solution approximates the user's intention in order to return better search results. Experiments show significant enhancement in search results over traditional keyword-based search engines' results and over selected semantic search engines.

1 INTRODUCTION

The problem of enhancing the search engine result has been tackled by many research in literature. The majority of this research (Surdeanu, et al., 2008) and (Verberne, et al., 2010) relies on question classification, linguistic pattern matching between questions and answers, focusing mainly on user query terms. Other research has introduced state-of-the-art ontology-based query/search systems as in (Maedche, et al., 2001); such systems help the user build a query-by-example, which by many users was found difficult. Personalized search solutions as in (Shirazi, et al., 2009) rely on availability of users' profiles in search sessions which limit the scope of the solution as sometimes the user may wish to explore relevant results without relying on any background or history he has.

Recently, a number of innovative semantic engine solutions have been introduced (Manuja and Garg, 2011). Despite the fact that Google is known as a keyword-based search engine, it has recently started injecting some semantic features to its technology (Allon, 2009).

We see today's search to be facing a two-fold problem — (1) lack of user ability to express his information need in accordance to how internet data is organized and (2) lack of reliable semantic-based search engines.

In this paper we propose a framework that consists of three main steps: (1) Enrichment of user query using a generic ontology, (2) Classification of user query into certain domain and (3) Reformulation of the enriched query using Domain-Specific Internet Data Organization Ontology (IDOO).

2 PROPOSED SOLUTION

Our solution bridges the gap between the user's lack of proper articulation of his information need and how the query should have been initially formulated to get a relevant and satisfying result. The framework of our solution is shown in Figure 1. We describe the details of the framework using the following running example that represents a shopping domain query:

Awad A., El-Sayed M. and El-Sonbaty Y..
APPROXIMATING USER'S INTENTION FOR SEARCH ENGINE QUERIES.
DOI: 10.5220/0003927604220425
In *Proceedings of the 8th International Conference on Web Information Systems and Technologies* (WEBIST-2012), pages 422-425
ISBN: 978-989-8565-08-2

Query: Where can I find a Cheap Kindle?

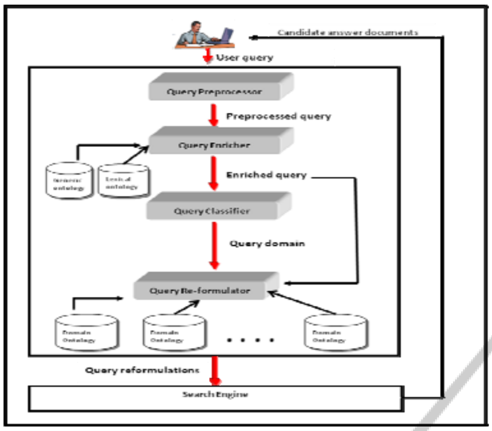

Figure 1: Proposed framework — ApproxMantic.

The first step in our solution is Query pre-processing. This step is mainly responsible for parsing the query and for performing term extraction, stemming and stop-word removal. When the query described above is given to the pre-processor the following terms are extracted: *Where*, *Find*, *Cheap*, *Kindle*. Next, the system semantically enriches the original query of the user using a generic ontology. Figure 2 shows part of a generic ontology that we use for our running example. The semantic enrichment is done by first mapping terms extracted from the original query, applying a spreading algorithm (Salton and Buckley, 1988) to activate other relevant nodes, and finally finding the shortest path among the activated nodes. This would normally introduce new concepts/nodes to the enriched query.

Figure 2 shows that the terms *Question, Investigation, Action, E-reader, Electronics, Product, Price, Criteria and Judgment Measure* are added to the enriched query as a result of the spreading. After applying the shortest path processing, the term *Location* is added to the enriched query.

The enriched query is passed next to a classifier to categorize it into a specific domain. We recommend the classifier presented in (Mostafa et al., 2009), since it is semantic based. The query enrichment step enhances the classification accuracy. For our running example, the query is classified to the shopping domain.

The final step in our solution, performed by the Query Re-formulator, takes as input the enriched query and the domain it has been classified to. The Query Re-formulator utilizes a special ontology that we propose, called IDOO (Internet-Data-Organization Ontology), for query reformulation. IDOO is domain-specific and is built by domain experts together with ontology engineers. IDOO models how domain-specific data is organized and expressed on the internet. We believe that the consideration of such valuable knowledge is very important to query reformulation.

IDOO of a domain mainly models knowledge of how terms and phrases of that domain are organized and presented in online documents related to that domain. It also models knowledge of keywords that are typically used to search this knowledge and their associations. IDOO mainly has concepts and relationships representing terms and phrases widely used in online documents related to the domain. It also includes Reformulation Rules (RR) for concepts, each rule is of the form: RR_i: $TT_i \rightarrow RT_i$ where TT_i is a triggering term and RT_i is a reformulating term. A triggering term represents a

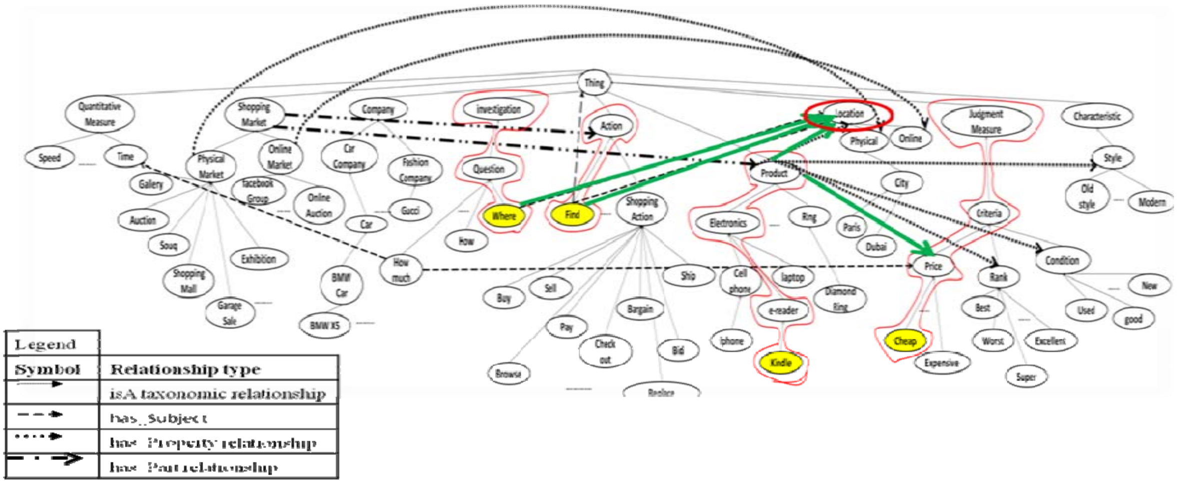

Figure 2: Query enrichment.

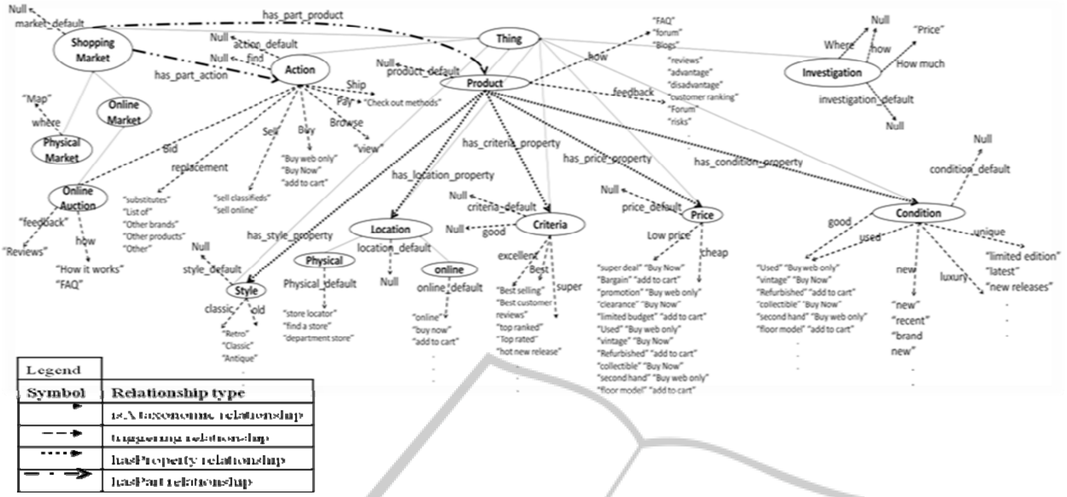

Figure 3: Part of the Shopping domain IDOO used in running example.

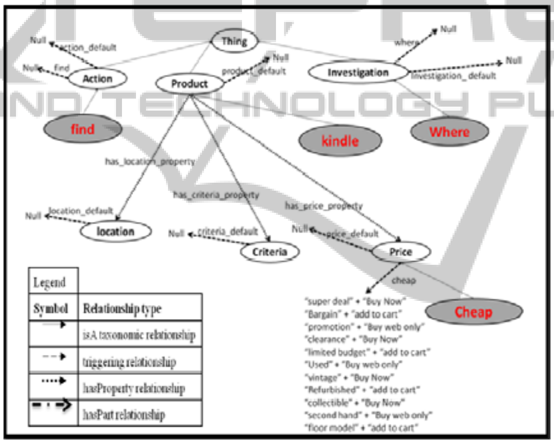

Figure 4: Mapping enriched terms to IDOO.

term that if exists in the original query, a proposed reformulating term or phrase is proposed in the new reformulated query. A concept in IDOO may have a default reformulation rule. The default reformulation rule is represented as: $RR_{default}$: default → RT_i. Figure 3 shows part of a Shopping domain IDOO that we use in our running example.

We use a special reformulation algorithm, described in more details in (Awad, 2012), to process the enriched query and to generate query reformulations. We now show how the algorithm works for our running example. Figure 4 shows the terms from the enriched query mapped to nodes on IDOO, particularly the terms *Action, Price, Criteria, Product, Investigation and Location.* The *Investigation* node has a reformulation rule: Where → Null triggered by the term "Where" in the original query, which means that this term should be removed from the reformulated query. Similarly, the *Action* node has a reformulation rule: *Find → Null* triggered by the term "find" in the original query, which also means that this term should be removed from the reformulated query. The *Price* node has reformulation rules: *Cheap → "super deal"+"buy now", Cheap → "bargain"+"add to cart"…etc,* triggered by the term "cheap" in the original query. The nodes *Action, Product, Criteria, Investigation, Location and Price* have default reformulation rules: *$RR_{default}$: default → Null,* which implies that these terms in nodes are removed from the reformulated query. The term "kindle" in original query that didn't map to any concept in IDOO and did not relate to any reformulation rule is kept in the reformulated query.

Next, the proposed reformulated queries are generated. Note that there might be several reformulations for each original query. Here are the generated reformulated queries:

Reformulations (some):

Kindle "super deal" "Buy now"
Kindle "bargain" "add to cart"
Kindle "floor model" "Buy web only"
…etc.

Finally the reformulated queries are passed to a search engine to retrieve results.

3 EXPERIMENTS

We assess the improvement in search results by observing the number of URLs returned in the result and the percentage of URLs in the top 20 URLs returned that are relevant to the user's query (top 20 results' precision).

We ran several categories of experiments. We investigated our solution with keyword-based search engines, such as Google and Yahoo. We also experimented against semantic-based search engines, such as Kngine and Hakia (Manuja and Garg, 2011).

The results obtained from our experiments show that our solution enhances both the results retrieved by the keyword-based search engines as well as by the semantic-based search engines. The best result achieved was when our solution was integrated with Google search engine. Figure 5, shows summary of top 20 result precision.

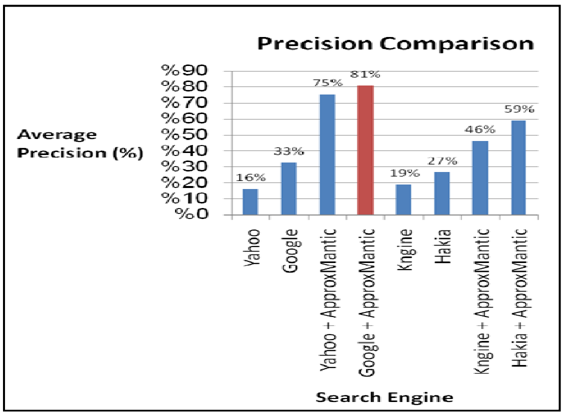

Figure 5: Summary of top 20 result precision.

4 CONCLUSIONS

In this paper we have introduced a novel technique for approximating search engine user's intention that relies on query reformulations. Our solution utilizes a special domain specific ontology that models how data on the internet is organized. Such ontology enables encoding and processing of query reformulation rules. Our experiments confirm that the proposed solution is able to generate highly relevant query transformations that deliver better results than that obtained from traditional term-based query engines and even that of selected semantic query engines.

REFERENCES

Allon, O. (2009). *Two new improvements to Google results pages.* Retrieved on December 7th, 2011. from http://googleblog.blogspot.com/2009/03/two new-improvements-to-google-results.html.

Awad, A. (2012). Approximating User's Intention for Search Engine Queries. M.Sc. Thesis. *College of Computing & I.T., Arab Academy for Science, Technology & Maritime Transport,* Alexandria, Egypt.

Maedche, A., Staab, S., Stojanovic, N., Studer, R., Sure, Y. (2001). SEAL - a framework for developing semantic web portals. *18th British National Conference on DB*, pages. 1–22.

Manuja, M. and Garg, D. (2011). Semantic Web Mining of Un-structured Data: Challenges and Opportunities. *International Journal of Engineering*, CSC Press, *Volume (5)*: Issue (3), pages 268-276.

Mostafa, L., Farouk, M., Fakhary, M. (2009). An Automated Approach for Web Page Classification. *19th International Conference on Computer Theory and Application*, Alexandria, Egypt.

Salton, G., and Buckley, C. 1988. On the use of spreading activation methods in automatic information. *11th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 147–160.

Shirazi, H. M., Shirazi, M. M. and Fardroo, N. (2009). Discovering User Interest by Ontology-based User Profile. *Int. Journal of Intelligent Information Technology application*, Engineering Technology Press. *Volume 2 [1]*.

Surdeanu, M., Ciaramita, M. and Zaragoza, H. (2008). Learning to Rank Answers on Large Online QA Collections. *ACL*, OH, pages 719-727.

Verberne, S., Boves, L., Oostdijk, N. and Coppen, P. (2010). What Is Not in the Bag of Words for Why-QA? *ACL*, pages 719-727.