# Query Reformulation Using Ontology and Keyword for Durian Web Search

Azilawati Azizan,
Nurazzah Abd Rahman
Faculty of Computer & Mathematical Sciences,
Universiti Teknologi MARA,
40450 Shah Alam, Selangor Malaysia
azila899@perak.uitm.edu.my

Zainab Abu Bakar
Al-Madinah International University,
40100 Shah Alam, Selangor, Malaysia
zainab.abubakar@mediu.edu.my

Shahrul Azman Noah
Faculty of Information Science & Technology,
Universiti Kebangsaan Malaysia
43600 Bangi, Selangor, Malaysia
samn@ukm.edu.my

***Abstract*—Query reformulation techniques based on ontological approach have been studied as a method to improve retrieval effectiveness. However, the evaluation of this techniques has primarily focused on comparing the technique with ontology and without ontology. The aim of this paper is to present, evaluate and compare the proposed technique in four different possibilities of reformulation. In this study we propose the combination of ontology terms and keywords from the query to reformulate new queries. The experimental result shows that reformulation using ontology terms alone has increases recall and decreases precision. However, better results were obtained when the ontology terms being combined with the query's keywords.**

***Index Terms*—query reformulation, ontology, query keyword, recall-precision, durian**

## I. Introduction

Most people think that searching information on the Web is easy as long as you have Google. However, Google or any search engine can be very helpful if the user can express and transform their information needed accordingly. In fact, the user faces a lot of problems to interpret and formulate the query correctly. That is why user reformulates query several times to get better results.

The size of information on the Web is increasing extremely that makes the searching activity become more complex. Even though there are Google, Bing or any search engine can assist user to find information, the effectiveness of those search engines really depend on the submitted query. The user will only receive relevant information if the submitted query precisely represents the user's information needs. That's why user has to specify their information needs correctly.

In reality, most users encounter a lot of difficulties to translate their information need into a query [1]. In addition, research in web query log also proved that many queries were badly formulated and nearly 50% of web user do reformulate their queries [2]. Looking at this problem, a lot of research and ideas have been suggested to assist users in formulating a better query. Among the approaches being suggested are query recommendations, query refinement [3],[4], query expansion [5],[6],[7], query disambiguation [8] and query reformulation [9],[10].

For each of the approach, there are many techniques being used by the researchers, such as query log [11],[12], rhetorical structure [13], thesaurus [14] and ontology [15],[16] . In the past years, the usage of ontology has attracted much attention in the information retrieval field. Researches have attempted to manipulate the use of ontology in various processes in the Information Retrieval (IR) model. This is because, ontology has proven to be very significant in improving search results especially for domain specific search [17],[18],[19].

Ontological approach in query reformulation is widely known as a method for improving the effectiveness of the retrieval system. However, the evaluation for this technique mostly focused only on presenting the comparison of retrieval effectiveness between ontology and without ontology technique. Looking at this lacking of evaluation view, we aim to present and compare the evaluation of one proposed technique with several different possibilities of reformulation.

This experiment is also an extension of our initial experiment done earlier [20] . We would like to identify which query reformulation techniques can give a more significant retrieval result than the previous work. In this paper we will discuss more on methods and analysis.

## II. Related Works

Query reformulation is a process of altering or modifying an initial query to improve search results and retrieval performance. It is an iterative process between users and search system in which user engage to find useful information that could satisfy their search goals. This is because effective query reformulation can improve the search results. It has been proved by works done by [21], [22] and [5].

In the early years of query reformulation research, Fidel [23] has identified three problematic situations that directed users to reformulate query: (1) retrieved sets were too large; (2) retrieved sets were too small; or (3) retrieved sets were off-target. Fidel also has classified query reformulation into two categories of query reformulation: operational and conceptual. In 1999, Lau and Horvitz [24] has developed a method to classify queries automatically into four exclusive types; generalization, new, reformulation and specialization. This

taxonomy was also used by many other researchers to detect search boundaries in search log automatically.

Then Rieh and Xie [22] extended the Fidel's [23] categories from two to four categories in 2006. They examined query reformulation by looking at the syntactic and semantic aspect. They found that, there are some common strategies being used by the users to perform query reformulation such as specification, generalization, replacement and parallel movement.

The study of query reformulation type has been expanded when Huang and Efthimiadis [25] published an extensive taxonomy consisting of twelve query reformulation types. Previous studies especially in identifying the categories of reformulations have inspired current researchers to explore more techniques in query reformulation field.

## III. MATERIALS AND METHODS

**Query Set:** Our query set contains approximately 290 queries. The queries are all domain-related (durian domain). The queries were gathered from online forums, blogs, social networks, online survey and Google instant feature. All the queries had undergone cleaning process to remove unrelated, incomplete and redundant queries. Table I shows some of the queries in our collection.

TABLE I. QUERIES

| Query No | Initial Tested Query |
|---|---|
| Q1 | List of insect pests that attack the durian tree |
| Q2 | When is the durian season in Malaysia |
| Q3 | What are the varieties of durian in Malaysia? |
| Q4 | What are the characteristic of good quality durian |
| Q5 | How to plant durian? |
| Q6 | How to control durian tree disease? |
| Q7 | What are the products of durian? |
| Q8 | What are the side effects of eating durian to health? |

**Search Engine**: Any search engine can be used to test our approach. In our case we chose the most popular search engine; Google to be our test bed.

**Ontology**: We employed DuriO (Durian Ontology) in our proposed technique. DuriO has been created in the earlier phase [26]. It has been created using Allegro Graph. All the facts in DuriO has been collected from trusted sources and it has been verified by durian experts from Malaysia Agriculture Research and Development Institute (MARDI). DuriO has more than 980 triples.

**Evaluation Metric**: For the purpose of evaluation, standard precision and average precision calculation was employed to measure the retrieval performance.

**Methods:** We proposed a technique that combine the query's keyword and the ontology terms. We execute four sets of experiment in order to compare the performance of the proposed technique. Firstly we submit the original query to the search engine as a benchmark. Secondly we reformulate the original query by just substituting it with keyword from the original query. Thirdly we reformulate it into ontology term alone. Lastly we combine the keyword and the ontology terms. Then we analyse the results using the precision and recall measures.

## IV. IMPLEMENTATION

This section briefly describes the implementation of the proposed technique. All the test query being submitted to the search engine accordingly based on the proposed technique. Then, twenty results were recorded for each of the submitted query. The results were examined and analyzed using the precision & recall calculation. Samples of retrieved results are shown in Fig 1. All initial queries (original query) are in natural language. The initial queries were reformulated into new queries according to the suggested techniques.

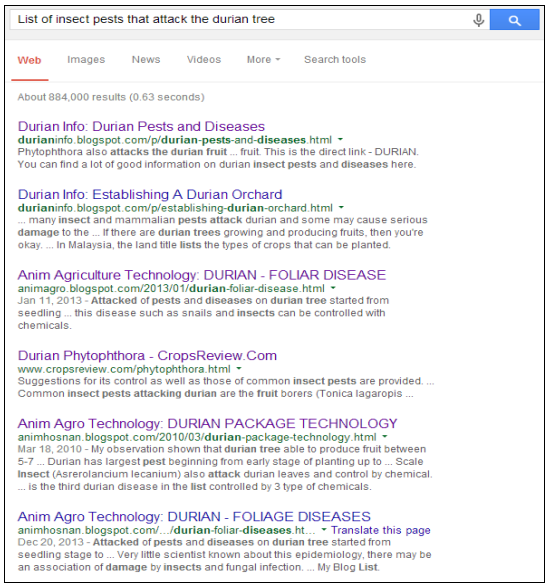


Fig. 1. Screen Shot of Retrieval Results

For the first technique, the original query was submitted to the search engine and then the retrieved results were captured. Example of query being submitted is "*List of insect pests that attack the durian tree*".

For the second technique, only keywords were submitted to the search engine. Based on the example above, keywords submitted were "*insect pest attack durian tree*". Those keywords were extracted from the original query using Stanford Parser [27]. Stanford Parser is a natural language parser. It is used to tokenize and tag words according to the part-of-speech tag in natural language processing field. Only words tagged as noun, adverb and adjective were chosen as the keyword.

**TABLE II.** REFORMULATED QUERIES

| Original Query (natural language) | Keyword | Ontology Terms (answers given by SPARQL) | Keyword + Ontology Terms |
|---|---|---|---|
| List of insect pests that attack the durian tree | Insect pest attack durian tree | fruit borers<br>mudaria<br>magniplaga<br>termite<br>microtermes<br>platypus beetle<br>teritip<br>pectenococus<br>koya (mealy bugs)<br>pseudococcus | Insect pest attack durian tree AND fruit borers<br>mudaria magniplaga<br>termite microtermes<br>platypus beetle<br>teritip pectenococus<br>koya (mealy bugs)<br>pseudococcus |
| When is the durian season in Malaysia | Durian season Malaysia | west coast peninsula – November, December , may, june<br>east coast peninsula – august and september | Durian season Malaysia AND west coast peninsula NovemberDecember may june,<br>east coast peninsula august september |
| What are the characteristics of a good durian fruits? | Characteristic good durian | no crack no hole strong aroma light color tough slimy less sharp large distance between spine | Characteristic good durian AND no crack no hole strong aroma light color tough slimy less sharp large distance between spine |
| List of durian varieties in Malaysia | Durian varieties Malaysia | masMuar<br>rajaKunyit<br>D99<br>D145<br>MDUR78<br>D189<br>MDUR88<br>Chanee<br>D190<br>D197<br>datoNina<br>D24<br>bukitMerah<br>musangKing<br>MDUR79<br>D168<br>D123 | Durian varieties Malaysia AND<br>masMuar<br>rajaKunyit<br>D99<br>D145<br>MDUR78<br>D189<br>MDUR88<br>Chanee<br>D190<br>D197<br>datoNina<br>D24<br>bukitMerah<br>musangKing<br>MDUR79<br>D168<br>D123 |
| How to plant durian | plant durian | field preparation, water irrigation, pruning, protection management, fertilizer, vegetative | plant durian AND field preparation, water irrigation, pruning, protection management, fertilizer, vegetative |
| What are the side effects of eating durian to health | Side effect durian health | diabetes<br>hypertension<br>blood pressure | Side effect durian health AND diabetes<br>hypertension<br>blood pressure |

For the third technique we substitute the original query with terms from ontology. Examples of the submitted terms were "*fruit borers, mudaria magniplaga, tonica terracella, termite microtermes, platypus beetle, mealy bugs*". This technique translated the original query into SPARQL query beforehand. Then it was used to query the domain ontology (DuriO) [20], [26]. Answers given by the ontology (ontology terms) was used to substitute the original query.

The fourth technique was a combination of the second and third technique. We combined the keyword from the original query with the ontology terms obtained. Examples of the reformulated query for this technique were "*insect pest attack durian tree* AND *fruit borers, mudaria magniplaga, tonica terracella, termite microtermes, platypus beetle, mealy bugs*".

We carried out the experiment by submitting the query one by one according to the query's number for all the suggested techniques. Then all the retrieved results were recorded accordingly. Samples of the reformulated queries according to the suggested technique are shown in Table II.

## V. RESULTS & DISCUSSION:

The popular IR evaluations metric is the precision and recall calculation. It is respectively formulated as in equation 1 and 2:

$$\textbf{Precision} = \frac{Number\ of\ relevant\ links\ retrieved}{Number\ of\ links\ retrieved} \quad (1)$$

$$\textbf{Recall} = \frac{Number\ of\ relevant\ links\ retrieved}{Number\ of\ relevant\ links\ in\ the\ collection} \quad (2)$$

In a normal controlled experiment, both precision and recall value can be easily calculated. We can exactly know the *number of relevant documents retrieved*, the *number of documents retrieved* and the *number of relevant documents in collection*. However for this experiment, we can only calculate and measure the precision value. The recall value can't be calculated since this experiment used Google as the test environment, which is impossible for us to know the exact number of relevant documents in the entire Google's collection.

Therefore, we logged the number of retrieved links for each query submission session according to the respected techniques to represent the recall. Although the total number of retrieved links given by Google is said to be only as estimation figure, but at least it can give us a general idea of the recall value for the given query. Table III exhibit the total and average of retrieved links.

TABLE III. NUMBER OF RETRIEVED LINKS

| | Natural Language | Keyword | Ontology Terms | Keyword + Ontology Terms |
|---|---|---|---|---|
| Q1 | 912,000 | 562,000 | 261,000 | 2,190 |
| Q2 | 321,000 | 325,000 | 462,000 | 178,000 |
| Q3 | 92,000 | 87,000 | 1,460,000 | 53,100 |
| Q4 | 129,000,000 | 271,000 | 21,400,000 | 5,310 |
| Q5 | 456,000 | 458,000 | 355,000,000 | 147,000 |
| Q6 | 167,000 | 112,000 | 8,150,000 | 52,900 |
| Q7 | 18,300,000 | 612,000 | 574,000 | 543,000 |
| Q8 | 241,000 | 3,160 | 310,000,000 | 2,590 |
| **Tot** | **149,489,000** | **2,430,160** | **697,307,000** | **984,090** |
| **Avg** | **18,686,125** | **303,770** | **87,163,375** | **123,011** |

In most query session, ontology terms techniques retrieved the highest number of links (represent recall), which is not good for retrieval performance. The most efficient retrieval system is to receive high precision and low recall. We can see a striking difference on the graph shown in Fig 2, the retrieved links for ontology terms techniques has increased steeply.

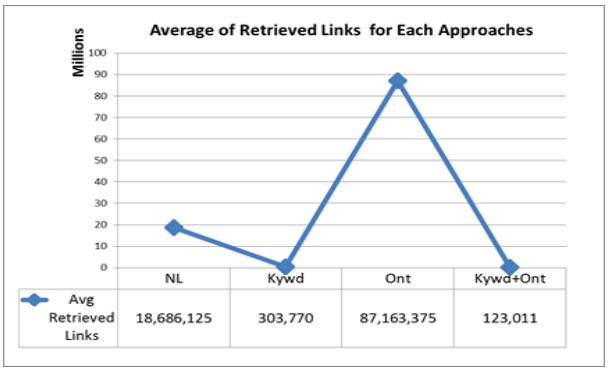


Fig. 2. Average of Retrieved Links

In order to measure and compare the quality of our techniques, we used the average precision calculation as in equation 3. In this comparison, we chose to measure at 10 cut-off point. It is common to measure the cut-off point for web retrieval performance either at 5, 10 or 20. We chose to measure at 10 cut off point since most users only inspect the first top ten of the links given in the retrieved result [28].

$$\bar{P}(r_j) = \sum_{i=1}^{Nq} \frac{P_i(r_j)}{N_q} \quad (3)$$

Fig 3 shows the average precision graph for the respected techniques.

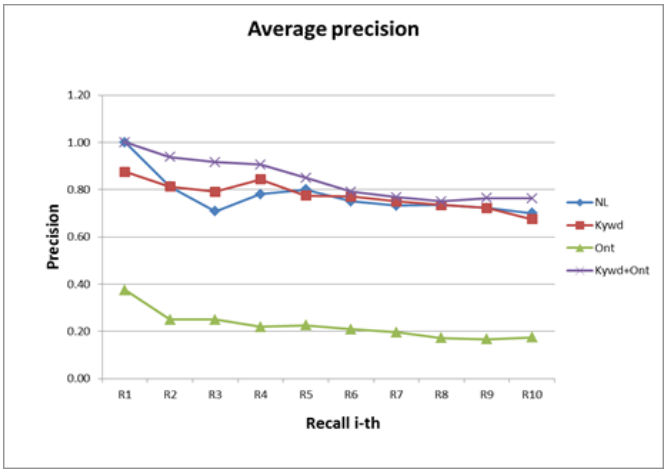


Fig. 3. Average Precision

The graph shows that Ont technique provides poor precision compared to other techniques. While Kywd+Ont has the highest precision value for almost all retrieved link at i-th.

This situation indicates that ontology terms alone are not enough to represent the user's query. It also points out that substitution of original query with ontology terms alone totally changed the original query intent. It hurts the retrieval performance.

To explain what has led to this negative result, we chose Q5 (*How to plant durian*?) as an example. This original query has been translated into SPARQL and then the ontology gave *field preparation*, *water irrigation*, *pruning*, *protection management*, *fertilizer*, *and vegetative* as the answers for durian planting stage. When these terms (ontology terms) were used to substitute the original query and resubmitted to the search engine, it returned a lot of irrelevant links. It is because the terms *field preparation*, *water irrigation*, *pruning*, *protection management*,..*etc* are not exclusive for the durian domain only, it overlaps with other plantation domain too.

## VI. CONCLUSION

The objective of this works is to identify better query reformulation techniques in retrieving better results for the specified domain. In this paper, experimental data and results for four different techniques were demonstrated.

This study focused on the use of domain specific ontology in reformulating new query. The results showed that combination of keywords and ontology terms provide better precision, compared to keyword and natural language techniques. Other highlight in this experiment is the use of ontology terms alone technique. It decreases the precision badly and increases the retrieved results. This really hurts the performance of retrieval.

More experiments need to be done to investigate the best techniques in reformulating query. We hope that this research will inspire further research in the query formulation and reformulation techniques.

ACKNOWLEDGMENTS

This research is based upon work supported by Ministry of Higher Education (Malaysia) under Long Term Research Grant (LRGS) Scheme (LRGS/TD/2011/UiTM/ICT/01) and Universiti Teknologi MARA (UiTM). The author also would like to acknowledge all contributors, MARDI staff and others who have helped and greatly assisted in the research.

REFERENCES

[1] J. Huang and E. N. Efthimiadis, "Analyzing and evaluating query reformulation strategies in web search logs," *Proceeding 18th ACM Conf. Inf. Knowl. Manag. - CIKM '09*, p. 77, 2009.

[2] V. Dang and W. B. Croft, "Query Reformulation Using Anchor Text," 2010.

[3] J. Guo, G. Xu, H. Li, and X. Cheng, "A Unified and Discriminative Model for Query Refinement Categories and Subject Descriptors."

[4] R. Suresh and B. C. Rout, "Advanced Approach In Query Refinement Using Refinement Filters From Knowledge Base," pp. 52–57, 2014.

[5] C. Carpineto and G. Romano, "A Survey of Automatic Query Expansion in Information Retrieval," *ACM Comput. Surv.*, vol. 44, no. 1, pp. 1–50, Jan. 2012.

[6] H. Imran and A. Sharan, "Thesaurus and Query Expansion," no. November 2009, pp. 89–97, 2009.

[7] R. Navigli, P. Velardi, D. Informatica, and R. La, "An Analysis of Ontology-based Query Expansion Strategies," in *Proc. Workshop Adaptive Text Extraction and Mining*, 2003.

[8] D. Yang, D.-R. Shen, G. Yu, Y. Kou, and T.-Z. Nie, "Query Intent Disambiguation of Keyword-Based Semantic Entity Search in Dataspaces," *J. Comput. Sci. Technol.*, vol. 28, no. 2, pp. 382–393, Mar. 2013.

[9] S. Haiduc, G. Bavota, A. Marcus, R. Oliveto, A. De Lucia, and T. Menzies, "Automatic Query Reformulations for Text Retrieval in Software Engineering," pp. 842–851, 2013.

[10] O. Asfari, B. L. Doan, Y. Bourda, and J. P. Sansonnet, "Personalized access to information by query reformulation based on the state of the current task and user profile," *3rd Int. Conf. Adv. Semant. Process. - SEMAPRO 2009*, pp. 113–116, 2009.

[11] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," *Curr. Trends Database Technol. - EDBT 2004 Work.*, vol. 3268, pp. 395–397, 2005.

[12] Z. Zhang and O. Nasraoui, "Mining Search Engine Query Logs for Query Recommendation," *Proc. 15th Int. Conf. World Wide Web*, pp. 1039–1040, 2006.

[13] S. Inam, M. Shoaib, F. Majeed, and M. I. Sharjeel, "Ontology Based Query Reformulation using Rhetorical Relations," vol. 9, no. 4, pp. 261–268, 2012.

[14] D. Buscaldi, D. Buscaldi, P. Rosso, P. Rosso, E. S. Arnal, and E. S. Arnal, "A WordNet-based Query Expansion method for Geographical Information Retrieval," *Geographical*, 2005.

[15] G. Solskinnsbakk, "Ontology-driven query reformulation in semantic search," no. June 2007, p. 99, 2007.

[16] J. Wu, I. Ilyas, and G. Weddell, "A Study of Ontology-based Query Expansion," in *Technical report CS-2011-04*, 2011.

[17] R. Alfred, C. K. On, P. Anthony, P. W. San, T. L. Im, L. C. Leong, and G. K. Soon, "Ontology-based Query Expansion for Supporting Information Retrieval in Agriculture," in *The 8th International Conference on Knowledge Management in Organizations Springer Proceedings in Complexity*, 2014, p. pp 299–311.

[18] S. M. Patil and D. M. Jadhav, "Semantic Information Retrieval Using Ontology and SPARQL for Cricket," vol. 4, no. 2, pp. 354–363, 2012.

[19] M. Fernández, D. Vallet, and P. Castells, "Automatic Annotation and Semantic Search from Prot Conference Item."

[20] A. Azizan, Z. A. Bakar, and S. A. Noah, "Analysis of Retrieval Result on Ontology-Based Query Reformulation," in *IEEE 2014 International Conference on Computer, Communication, and Control Technology (I4CT 2014)*, 2014, no. I4ct, pp. 244–248.

[21] B. J. Jansen, D. L. Booth, and A. Spink, "Patterns of Query Reformulation During Web Searching," vol. 60, no. 7, pp. 1358–1371, 2009.

[22] S. Y. Rieh and H. (Iris) Xie, "Analysis of multiple query reformulations on the web: The interactive information retrieval context," *Inf. Process. Manag.*, vol. 42, no. 3, pp. 751–768, May 2006.

[23] R. Fidel, "Moves in Online Searching," *Online Rev.*, vol. 9 (1), pp. 61–74, 1985.

[24] T. Lau and E. Horvitz, "Patterns of Search:Analyzing and Modeling Web Query Refinement," *Proc. User Model. Conf.*, pp. 119–128, 1999.

[25] E. N. Efthimiadis, J. Huang, A. Spink, and J. Jansen, "Query formulation in web search," *Proc. ASIST Annu. Meet.*, vol. 46, 2009.

[26] Z. A. Bakar and K. N. Ismail, "Base Durian Ontology Development Using Modified Methodology," in *M-CAIT 2013*, 2013, pp. 206–218.

[27] "Stanford Parser." [Online]. Available: http://nlp.stanford.edu:8080/parser/. [Accessed: 20-Sep-2014].

[28] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. ACM Press. Addison Wesley, 1999.