Available online at www.globalilluminators.org

GlobalIlluminators

FULL PAPER PROCEEDING
Multidisciplinary Studies

Full Paper Proceeding ITMAR -2014, Vol. 1, 157-164

ISBN: 978-969-9948-24-4

ITMAR-14

Query Reformulation Using Domain Ontology

Azilawati Azizan [1*], Zainab Abu Bakar [2], Nor Adzlan Jamaludin [3] , Khairul Nurmazianna Ismail [4] & Rohana Ismail[5]

Faculty of Computer and Mathematical Sciences Universiti Teknologi MARA,Shah Alam, Selangor Darul Ehsan, Malaysia

Abstract

Indeed, today's search engines have made a giant leap towards helping users in finding relevant information within a large body of poorly organized information (i.e., the Web). However, existing search engine can perform well in retrieving relevant results only if the user submits a query that clearly reflects their information needs. But most users are unable to provide such queries. Motivated by this issue, this paper discusses an ongoing research project in developing a search engine using several elements of semantic technology. We propose to apply the domain ontology at the query processor module. Our approach proposed to reformulate the user’s natural language query by performing a SPARQL (Protocol and RDF Query Language) against the domain ontology. Then, the results from ontology were matched with the index file generated using conventional method. The ontology is assisting infinding answers and links to related documents and web sites. In this sense, user will retrieve more relevant information since the query has been refined with the knowledge in the domain ontology.

Keywords*:* Query Reformulation, Search Engine, Ontology, SPARQL, Natural Language

Introduction

At present day, finding information on the Web is not an easy task, since the Web content is unorganized and it is growing without limit. Search engine has tried hard in assisting the user to fulfil the user’s needs. And today’s search engine seems has done a good job in helping the user to dig out information from a huge unorganized library (Web). But actually the search engine is just helping us in providing tons of links (Ramachandran and Sujatha 2011) for users to inspect one by one in order to filter which link is actually providing the answer to the user.

We also cannot deny the capabilities of the current search engines such as Google and Bing are really effective in finding documents on the Web. Most of the times they return correct and relevant results to the user, if only when the user submits with correct query.

Usually user would typically prefer to submit query in their native language (Wu 2011; and Manning 2008) but this resulted in less relevant of searched result. It is because of the difficulties in translating the user query by the search engine. Then user attempts to

*All correspondence related to this article should be directed to Azilawati Azizan, Faculty of Computer and Mathematical Sciences Universiti Teknologi MARA,Shah Alam, Selangor Darul Ehsan, Malaysia

Email: zainab@tmsk.uitm.edu.my

reformulate query into a simpler and shorter query (Imran, H., & Sharan, 2009). The user also needs to reformulate queries several times for getting other related information for the same concept (Sharef 2012; and Abu Bakar 2012). This only makes the search result too general and high in recall.

In this project we are trying to improve the retrieval result by applying query reformulation on the user's query. We use domain ontology to assist the task of finding relevant answer based on the user's query. Instead of matching user's query terms with the index terms, we use the answer (terms) supplied by the ontology to match with the index's terms.

Related Work

Query reformulation is the process of modifying a query with the objective to improve retrieval results. The reformulations that can be performed on the query can include spelling suggestions and corrections, automated term suggestions, suggesting popular destinations, relevance feedback, and showing related articles (Hearst 2009). In most cases , query reformulation is performed with the support of a dictionary , query logs (Huang and Efthimiadis 2009) or ontology (Stevens et al. 2003; Bechara 2009; and Wu 2011).

Current query mechanisms are highly dependent on a prior understanding of the data model in the ontology. To query the ontology, users need to express their information needs in a graph query language such as SPARQL which is structured and formal representations of the relationships between the data. Natural language query is able to provide users for querying the ontology and thus ignore the complexity of formulating a query expressed in a SPARQL (Pradel et al. 2013).

Natural Language (NL) Query to SPARQL Tools

PowerAqua et al (2011), is a Question Answering (QA) system which able to answer queries by discovering and integrating information from distributed semantic resources. The system process and extract query using Gate NL processing tool to identify accurate queries formed with question wh-terms (which, what, who, when, where) or commands (give, list, show, tell, etc) with variation of length and complexity.

FREyA Damljanovic et al. (2012), is another QA system which employs an interactive Natural Language Interface for querying ontologies. It uses syntactic parsing in combination with the ontology-based lookup in order to interpret the question. The system involves user's feedback and clarification dialogs to resolve ambiguities in a query.

AutoSparQL Lehmann and Bühmann (2011), is also a QA system which can convert a natural language question to a SPARQL query for retrieving the answer of a question from a triple store. The system uses supervised machine learning and allows users to query the triple store without knowing the schema of the underlying knowledge base and without expertise in the SPARQL query language. The user interface has implements an active learning approach to support users.

Ontology

Since Semantic Web emerged, ontology has greatly being used to support Information Retrieval. Various aspect of ontology activities is being manipulate to improve the retrieval (Azizan et al. 2013). There are various definitions of ontology in Computer Science. One of the most widely accepted is the definition by Gruber, “An explicit specification of a conceptualization” (Gruber 1993).

Ontology in Computer Science evolved from semantic networks Studer et al. (1998), and extended to the knowledge sharing. It was recognize to be useful in representing and facilitating the sharing of knowledge about a domain by human and automatic agents. Ontology has been used in Configuration System, Software Engineering, Information Retrieval, Conceptual Modeling, Interoperability, Enterprise Modeling, Electronic Commerce, and many other fields in the research and production areas. Ontology has been adopted in various ways in Information Retrieval (IR) for achieving better retrieval results. For example, they employ ontology in the indexing process (Kara et al. 2012). Some employ it at ranking process (Aleman-Meza et al. 2010), and they also employ it for query expansion (Navigli and Velardi 2003; and Wu et al. 2011) and reformulation purposes.The research that related toontologyin thefield of IR isclosely related to theSemanticWeb. Ontologyis one of the componentinSemanticWebtechnology (Antoniou et al. 2005). Many researches claims that the existence of ontology in the Semantic Web can purposely producerelevant answers (Mukhopadhyay 2011).

Methodology

Main focus of this research is on the query formulation. We applied the SPARQL towardsthe domain ontology to reformulate the user query. Figure 1 shows the general architecture of our approach.

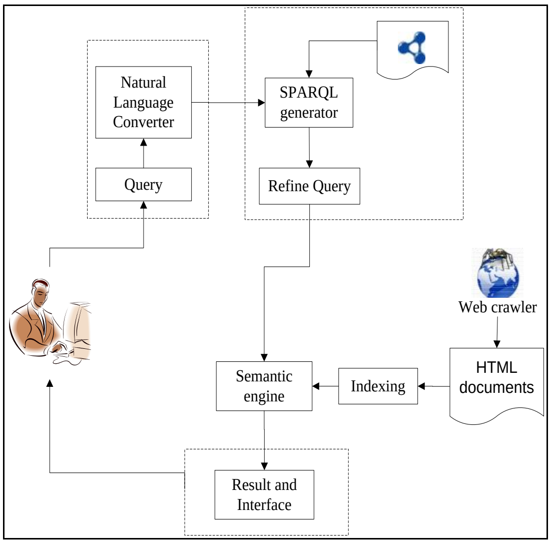


Figure 1: General Architecture of Our Approach

Conversion of Natural Language Query to SPARQL

Figure 2 illustrate our methods for converting from the user natural language into SPARQL query.

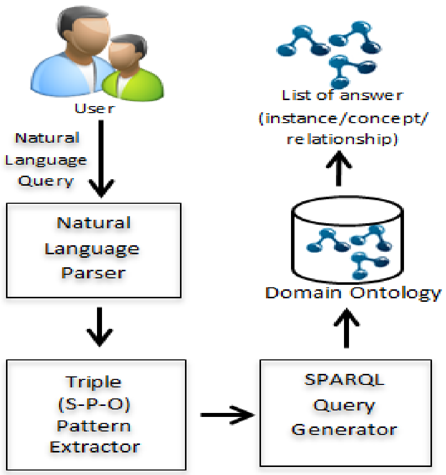


Figure 2: Conversion Natural Language to SPARQL

We perform Part of Speech (POS) tagging on the user's query. Afterwards, we attempt to identify patterns in the query in the form of subject-predicate-object (S-P-O) term(s) from the resulting tags. This S-P-O pattern isused as the basis for the SPARQL query generation.

From preliminary analysis of the 20 validated queries, we found that these queries tend to take certain patterns. As such, we created several SPARQL templates that reflect the identified patterns. During conversion, any terms contained in S-P-O pattern will first be lemmatized. After the lemmatization process is complete, we attempt to match the S-P-O patterns with the SPARQL templates. With this, we generate the SPARQL query that will be run against the ontology.

Indexing Process

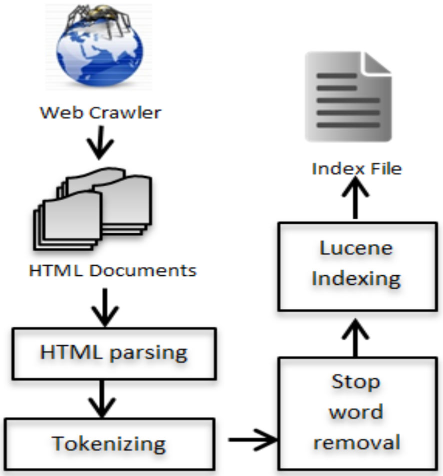


Figure 3 : Document Processing and Indexing

Indexing is a pre-process task. In this process, it begins with parsing the HTML documents.2000 HTML documents (from our data collection) are parsed to remove the HTML tags. Then all the terms are tokenized, and stop words are removed. After all, the inverted index is created and ready for use in the matching process. Figure 3 demonstrate how indexing process is done in our project.

Term Matching

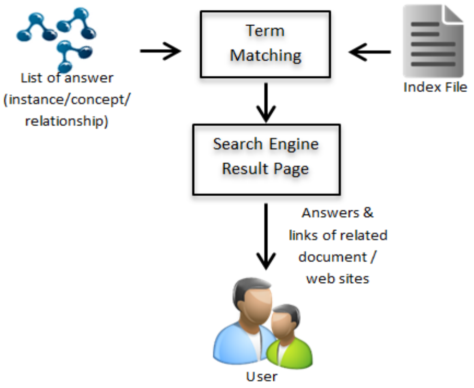

List of answer (instance/concept/ relationship)
Term Matching
Index File
Search Engine Result Page
Answers & links of related document / web sites
User

Figure 4: Retrieval Method

In a normal search engine, the retrieval part starts with the input query from the user. Then the user's query will be translated into a query terms or into a same structure of the index terms so that the matching process will be easier. In our approach the input query is provided by the answers given from the ontology. Results given from ontology could be in terms of concepts / instances / relationships in the domain ontology. It is then matched with the terms in the inverted index created beforehand. Then all HTML document matched with the ontology result is being retrieved and presented to the user. Figure 4 represent the above task.

The HTML result of this search engine will be presented to the user according to the results from the ontology. These organized results will ease the users to look for the expected relevance documents to the submitted query instead of mixing all the HTML documents together.

Implementation

Our test data consist of 2000 HTML documents that have been collected by our own crawler. Data cleaning is done manually. All non-related document, broken link documents, and duplicate content from different sources was removed to make sure the collection only contain quality documents. For initial testing, we only processed English documents.

We also provide query collection in our test data. We have collected hundreds of query that is related to our scope of domain. After doing some cleaning and grouping process, and also get validation from the domain expert, we published 20 queries that is the most common question being asked in the scope of the domain. All the queries are in natural language form. Figure 5 show example of the query collection.

In this project we modified the domain ontology from our previous work (Bakar and Ismail 2013). This is because the previous ontology primarily focuses on one aspect of the domain. We have enriched the ontology with additional classes, relations and instances to reflect broader knowledge of the domain.

What are the durian varieties?
What are the durian tree diseases?

Figure 5: Sample of NL Queries

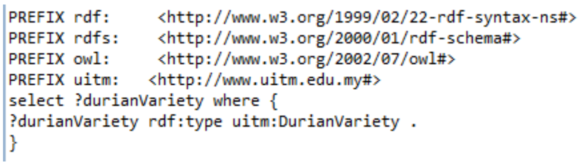
Figure 5 shows a portion of NL queries that we have collected. Figure 6 shows the POS tagging result for the query “What are the durian varieties?”. With the tagging at figure 6 we are able to extract the S-P-O pattern as shown in figure 7. The SPARQL generated from the S-P-O pattern is as shown in figure 8 and a sample of the results of the SPARQL query when run against the domain ontology can be seen in figure 9.

What\WD are/VBP the/DT

Figure 6: POS Tagging Result

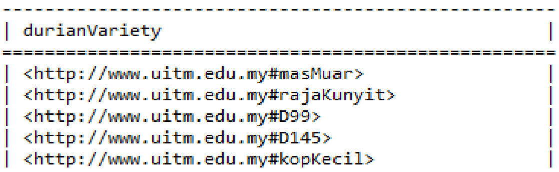
Subject Predicate Object
? typedurian varieties

Figure 7: S-P-O Pattern



```
PREFIX rdf:     <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:    <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl:     <http://www.w3.org/2002/07/owl#>
PREFIX uitm:   <http://www.uitm.edu.my#>
select ?durianVariety where {
?durianVariety rdf:type uitm:DurianVariety .
}
```

Figure 8: Sample of Generated SPARQL



```
---------------------------------------------------
| durianVariety                                   |
===================================================
| <http://www.uitm.edu.my#masMuar>                |
| <http://www.uitm.edu.my#rajaKunyit>             |
| <http://www.uitm.edu.my#D99>                    |
| <http://www.uitm.edu.my#D145>                   |
| <http://www.uitm.edu.my#kopKecil>               |
```

Figure 9: Sample of Results from Ontology

Conclusion

We believe that it is too early to conclude that ontology-based query reformulation is promising in improving the retrieval precision. Our approach suggests that by translating the user query into SPARQL and using it to manipulate the ontology will probably increase the relevancy. At current stage, we are revising our initial experiment towards a better ontology-based query reformulation strategy. Our future works will also focus on the search result presentation. We will organized the information given by the search engine according to the ‘WH’ question such as ‘what’, ‘why’, ‘when’, ‘who’, ‘where’, and ‘how’.

Acknowledgments

This research is supported by Ministry of Higher Education (Malaysia) under Long Term Research Grant (LRGS) Scheme (LRGS/TD/2011/UiTM/ICT/01) and UniversitiTeknologi MARA (UiTM).

References

Abu Bakar, Z., Ismail, K. N., & Fooladi, N. (2012, December). Effectiveness of query formulation based on durian characteristics. In *Research and Development (SCOReD), 2012 IEEE Student Conference on* (pp. 59-62). IEEE.

Aleman-Meza, B., Arpinar, B., Nural, M. V., & Sheth, A. P. (2010, September). Ranking documents semantically using ontological relationships. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on* (pp. 299-304). IEEE.

Antoniou, G., Christophides, V., Plexousakis, D., & Doerr, M. (2005). Semantic Web Fundamentals

Azizan, A., Bakar, Z. A., Khairuddin, N., & Saad, N. L. (2013). Ontology-Based Information Retrieval: A Review. In *International Symposium on Mathematical Sciences and Computing Research*.

Bakar, Z. A., & Ismail, K. N. (2013). Base Durian Ontology Development Using Modified Methodology. In *Soft Computing Applications and Intelligent Systems*(pp. 206-218). Springer Berlin Heidelberg.

Bechara, A., Campos, M. L. M., & Braganholo, V. (2009). Applying biomedical ontologies on semantic query expansion. *ICBO*, 158.

Damljanovic, D., Agatonovic, M., & Cunningham, H. (2012, January). FREyA: An interactive way of querying Linked Data using natural language. In *The Semantic Web: ESWC 2011 Workshops* (pp. 125-138). Springer Berlin Heidelberg.

Gruber, T. R. (1993). A translation approach to portable ontology specifications.*Knowledge acquisition*, *5*(2), 199-220.

Hearst, M. (2009). *Search user interfaces*. Cambridge University Press

Huang, J., & Efthimiadis, E. N. (2009, November). Analyzing and evaluating query reformulation strategies in web search logs. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 77-86). ACM.

Imran, H., & Sharan, A. (2009). Thesaurus and query expansion. *International journal of computer science & information Technology (IJCSIT)*, *1*(2), 89-97.

Kara, S., Alan, Ö., Sabuncu, O., Akpınar, S., Cicekli, N. K., & Alpaslan, F. N. (2012). An ontology-based retrieval system using semantic indexing.*Information Systems*, *37*(4), 294-305.

Lehmann, J., & Bühmann, L. (2011). Autosparql: Let users query your knowledge base. In *The Semantic Web: Research and Applications* (pp. 63-79). Springer Berlin Heidelberg.

Lopez, V., Fernández, M., Motta, E., & Stieler, N. (2011). Poweraqua: Supporting users in querying and exploring the semantic web. *Semantic Web*,*3*(3), 249-265.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1, p. 496). Cambridge: Cambridge university press.

Mukhopadhyay, D., Banik, A., Mukherjee, S., Bhattacharya, J., & Kim, Y. C. (2011). A domain specific ontology based semantic web search engine. *arXiv preprint arXiv:1102.0695*.

Navigli, R., & Velardi, P. (2003, September). An analysis of ontology-based query expansion strategies. In *Proceedings of the 14th European Conference on Machine Learning, Workshop on Adaptive Text Extraction and Mining, Cavtat-Dubrovnik, Croatia* (pp. 42-49).

Pradel, C., Haemmerlé, O., & Hernandez, N. (2013). Natural language query interpretation into SPARQL using patterns. In *Fourth International Workshop on Consuming Linked Data-COLD 2013* (pp. pp-1).

Ramachandran, A., & Sujatha, R. (2011). Semantic search engine: A survey.*International Journal of Computer Technology and Applications*, *2*(6).

Sharef, N. M., & Noah, S. A. M. (2012, December). Semantic search processing in natural language interface. In *Computing and Convergence Technology (ICCCT), 2012 7th International Conference on* (pp. 1436-1442). IEEE.

Stevens, R., Goble, C., Paton, N. W., Bechhofer, S., Ng, G., Baker, P., & Brass, A. (2003). Complex query formulation over diverse information sources in TAMBIS. *Bioinformatics: Managing Scientific Data*. *Morgan Kaufmann*.

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: principles and methods. *Data & knowledge engineering*, *25*(1), 161-197.

Wu, J., Ilyas, I., & Weddell, G. (2011). *A study of ontology-based query expansion*. Technical report CS-2011-04, University of Waterloo.