Research Journal of Applied Sciences, Engineering and Technology 11(5): 468-472, 2015
DOI: 10.19026/rjaset.11.1849
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: September 07, 2014 Accepted: October 17, 2014 Published: October 15, 2015

Research Article

Semantic Triple Ranking based on Levenshtien Reverse Engineering Approach

Aliyu Rufai Yauri, Rabiah Abdul Kadir, Azreen Azman and Masrah Azrifah Azmi Murad
Faculty of Computer Science and Information Technologi, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

Abstract: In sematic Web data are represented in Resource Description Framework (RDF) in triple format (Subject, relation, Object) and retrieved using structured query such as SPARQL. These structured queries require complex syntax to formulate. In view of this therefore, several approaches have been researched to enables semantic formulation of natural language to structure query. The process involves the representation of natural language query to structured triple format. However, dues complex nature of natural language, one natural language query may have more than one possible triple format; therefore an effective semantic triple ranking framework is needed for semantic triple ranking. In this study, semantic triple ranking mechanism is proposed. The approach is based on using levenshtien string matching algorithm a reverse engineering approach. The result of the proposed triple ranking has increased precision to 0.04 and recall 0.06.

Keywords: Concept, information retrieval, predicate, Quran ontology, semantic web, triple

INTRODUCTION

The volume of information online over the years has increased at exponential rate. This has increased the popularity of web search. However, current search engines such as Google and Yahoo are based on traditional keyword matching (Lixin and Guihai, 2006). Data retrieval in traditional key word matching lacks enough semantics which led to the retrieval irrelevant information. Current search engines have no means for specifying the semantics of data which makes it difficult for computers to understand the meaning for processing (Anna, 2012).

Semantic Web was introduced by W3C consortium to incorporate semantics into the search process. In semantic Web data is provided with more descriptions for machines to understand and process. In semantic Web Data are represented semantically into structured Resource Description Format (RDF) which enable representation of data into ontology concept linked with explicit relationships (Sreeja *et al*., 2012). Simple definition of ontology can be seen as any objects we can identify from domain and relationship that exist between those objects. Objects are mainly referred as concepts in ontology. RDF represents ontology into graphical triple form (subject, predicate, Object). Subject and objects are ontology concepts, predicate stands for explicit relationships that exist between Subject and Object. Predicate may be represented by a word, phrase or sentence. This semantically structurally represented RDF data is stored in the knowledge base in order to facilitate access and process. Knowledge base is storage fertility for semantically represented data.

For accessing data stored in the knowledge base, query has to be semantically formulated in the same representation with RDF triple format in the knowledgebase. However, due to complex nature of natural language, a given natural language query may be semantically formulated into more than one possible triple representation. In this case processing triple that is not the most appropriate triple representation of the query will affect the precision and recall of the retrieved information.

This study presented automated semantic triple ranking approach. The approach semantically rank triples in order to get the most likely triple representation of the natural language query from possible representations. The proposed triple ranking approach is based on using levenshtien string matching algorithm a reverse engineering approach. The knowledge base use for experiment involves the re-use of existing published by Leeds University United Kingdom. Leeds University ontology is built from Quran domain ontology which identified 300 important Noun concepts from Quran and 350 relationships. For this research, we updated Leeds Quran ontology by annotating the ontology using various Islamic related documents to increase the relationships to 1600 relationships. The Ontology we are experimenting contains only Noun concepts. Therefore focus of the Research it to enable user ask Natural Language Query concerning noun concepts in the Quran. We will see in details in the remaining part of this study.

Corresponding Author: Aliyu Rufai Yauri, Faculty of Computer Science and Information Technologi, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

LITERATURE REVIEW

The emergence of semantic web technology, semantics has attracts the interest of current major search engines, such as Yahoo, Search-Monkey, Google and other search engine (Ivan and Miloslav, 2013). This has made the exploration of semantic Web content and the amount of linked data increased (Vanessa *et al*., 2010). Despite the availability of linked data today, the task of supporting user to retrieve these linked data remain a great challenge. Researches have been on, on how best user is supported in order to retrieve important knowledge from these linked data. Since semantic web technology frame work is based on triple representation of RDF data and user Query is Natural Language, user query need to be transformed into triple format and semantically ranked the triples in order to have access to these linked data in more effective manner. Triple ranking is the process of ranking semantically formulated triples in order to get the closest triple representation of natural language query ranked with highest weight.

Several researches has been presented ranking of resources based on keyword input such as Swoogle (Tim *et al*., 2005) and ReConRank (Aidan *et al*., 2006). Relying on occurrence of keyword in order to rank triple will eliminate semantic into the retrieval process and as a result wrong triple may be presented with highest weight even though semantically it is not the best triple representation of the natural language query. Furthermore, (Vanessa *et al*., 2005) presented an approach that rank triple relations based on associate tf-idf, which is used to find related concepts in the ontology given an initial set of concepts and corresponding initial activation values. (Kemafor and Amit, 2002) presented an approach that attempt to find semantic similarity between paths connecting different triples in RDF model. Research in Ramakrishnan *et al*. (2005) proposes a heuristic method for weighting graph patterns connecting two nodes in a graph considering the differences of edges given by RDF graph that includes schema information encoded as RDFS ontologies. These approaches main objective is to measure similarity and rank triples with the knowledge base. Triple within the knowledge base are already structured, dealing unstructured natural language query will be a different. Where each of triples that may be formulated from the natural language query may all be similar to particular triple in the knowledge base. Therefore the task of ranking triple formulated from natural language query using these methods may not be easy.

Some recent work has focuses on ranking triple formulated from natural language query. Works in Damljanovic *et al*. (2012) and Franz *et al*. (2009) ranked triples that were formulated from natural language by ranking the relationship detected from the natural language query. However, ranking relations may be easier in the case where the query has one or two concepts identified. Or in the case where user is involved in triple formulation. Where relations are ranked and are presented to the user to map with the identified concepts in order to generate triple. And this may require user going through a lot of process by mapping the relations with concepts in the case where the queries has many concepts.

This study proposes a triple ranking mechanism that focuses on ranking the complete triple (subject, relation, Object) instead of just ranking relation. The system focuses on raking the automatically generated triple from natural language. We argue that, ranking the entire triple instead of relation will give more effective semantic triple ranking.

METHODOLOGY

Triple ranking approach proposed in this study is based on using levenshtien string matching algorithm a reverse engineering approach. Figure 1 shows flow chart of the proposed triple ranking system.

Figure 1 presents the flowchart of the proposed triple ranking system. The system accepts natural language query and attempt to semantically formulate the query into structured triple representation. The natural language query goes through normalization. The normalized natural language query is used to semantically formulate the query to triple representation of the query. The semantic query formulation is done using statistical machine learning approach to automatically identify concepts from the query and automatically detect possible predicates between the identified concepts. Concept identification involved automatically matching the noun token of the query against the ontology gazetteer. If any of the noun query tokens match any concept in the gazetteer, such token is automatically identified as concept. The next task is to use the remaining query tokens to automatically detect phrase or sentence as the possible predicate between the earlier identified concepts. The task of predicate detection involves learning from the knowledgebase to automatically predict phrase or sentence as possible predicate by using Ngram maximum likelihood estimation. Ngram is an automata for predicting phrase or sentence by estimating the probability of a word given the previous word. Since the focus of this paper is about triple ranking, we focused on showing the details of triple ranking in this study.

When the system semantically formulates natural language query to structured triple representation, due to complex nature of natural language, more than one possible triple representations of the query may emerge. Semantic Triple ranking is required to rank these triple in order to retrieve relevant information using the most ranked triple. In this study the ranking method is based

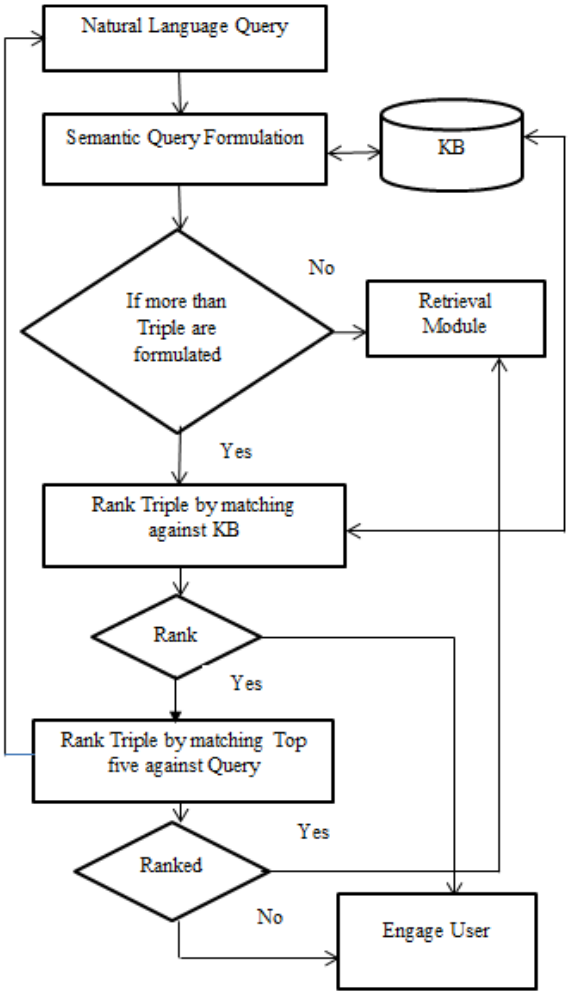

Fig. 1: Flowchart of the semantic triple ranking

on using Levenshtien string matching algorithm, a reverse engineering approach. Levenshtein string similarity algorithms compute the distance between two given strings. They shows the minimum number of operations that are needed to transmute one string into another. The operations are comprised of insertions, deletions, or substitutions of a single character. Levenshtien computation is normalized and assigned a score in the range 0 and 1. For example, using the Levenshtein distance between "kitten" and "sitting" is 3, i.e., the number of operations comprising the substitution and insertion needed to convert kitten to sitting is 3. The application of levenshtien string matching algorithm for ranking proposed in Damljanovic *et al*. (2012). In their work they used Levenshtien string distance rank predicates that were detected by their system. In this study, in order to get the most suitable triple representation of the query when more than one possible predicate were detected from a query and thus more than one triple representation of the query is formulated by the system. Reverse engineering method for triple ranking in this

So many prophets have been reported to have been sent by God to the world. Who is the last prophet among them and how can you prove that he is a prophet?

Fig. 2: Sample query

(?, is the last, Prophet) and (?, is-a, Prophet)

Fig. 3: Formulated triple

study is implemented by using the Levenshtien string matching algorithm to match the initial top five ranked triples against the original queries which resulted in obtaining the most appropriate triple representation of the queries being ranked with highest score.

In the proposed method, the system takes automatically formulated triples as input and perform initial ranking by performing string matching using levenshtien string matching algorithm against triples in the knowledge base. In a reverse engineering method, top 5 initially ranked triples are then matched against the natural language query also using levenshtien string matching algorithm. The triple that has the highest weight is then assume to be the closet triple representation of the given natural language query. Figure 2 show example of natural language query used to show the implementation of proposed triple ranking in this study.

Figure 2 present natural language query. Based on the semantic query formulation algorithm, the query generated 2 possible triples. The natural language query in Fig. 2 is semantically formulated into two possible triple as seen in Fig. 3.

Figure 3 presents possible triple representation of the natural language query in Fig. 2. The triple are formulated in based on triples in the knowledge base which the algorithm used for predicate prediction. In this case, the triple in Fig. 3 exist in knowledgebase, therefore the algorithm is able to use these triple in the knowledge base and automatically formulate the query to triples in Fig. 3. In this case, since the query produce more than one triple, processing any of the triple without ranking may end up retrieving irrelevant result. Since more than one triple are formulated from the query, the system automatically parse the formulated triples to the ranking approach for ranking.

The ranking approach, takes the automatically formulated triples representation of the natural language as input. The triple ranking process starts by matching the formulated triple against the knowledge base using a Levenshtien string matching algorithm. The system automatically takes the first 5 ranked triple in case the triple used for the ranking are >5. In the case of example in this study the formulated triple are 2. So the two triple are ranking. The system then uses the automatically ranked triples and performs a reverse

engineering approach against the natural language query. Based on the experiment, although the initial ranking ranked and reduce the number of possible triple query representation to 5 in the case where more than five triple representation of the query are formulated, the reverse engineering prove to get better triple representation of the query. The returned results of the reverse engineering approach prove to be more relevant than the initial ranking.

For example, from the triples representation of the query in Fig. 3, after the ranking returned, ?, is-a, Prophet with higher weight than ?, is the last, Prophet. However, is-a, Prophet is not the most appropriate triple representation of the query in Fig. 2 because we are comparing it with any of the triples in the training set. Let's assume that the triples in the training set are Muhammad, is the last, Prophet and Isah, is-a, Prophet and we are comparing with these with the formulated triples. Because Muhammad has more characters than Isah, ?, is-a, Prophet will have highest score than ?, is the last, Prophet. However, although, ?, is-a, Prophet has a higher weight it is most likely that ?, is the last, Prophet is closer to a triple representation of what the user is trying to search for. Therefore in order to obtain the triple that is closer to the query words, this paper employed a reverse engineering approach by computing the distance between the ranked triples against the query, i.e., the minimum distance in transforming any of the ranked triples to the original query. This approach gave better results in terms of obtaining the triple representation most appropriate to the triple representation with the highest score.

For example, after applying the reverse engineering approach, ?, is the last, Prophet has a higher score and thus is accepted by the system as the most appropriate triple representation of the natural language query in Fig. 2. The ranked triple is the parsed to the retrieval module for retrieval of the relevant I formation.

In the case the system is not able to automatically rank the formulated triple; user is engaged to choose from the formulated triples. The triple that is the chosen by the user is then used by the system for retrieval of the relevant information.

In the next section, an evaluation and analysis of the semantic ranking approach is presented.

RESULTS AND DISCUSSION

In this section, evaluations of effectiveness proposed triple ranking method is presented. The effectiveness of the proposed triple ranking method was measured based on improvement of the precision and recall of the retrieved result. For experiment, the ranked triple is used to semantically retrieved answers from the knowledge base, where the precision and recalled of the retrieved answer by the proposed method is compared with relation ranking approach in FREyA (Damljanovic *et al*., 2012). Table 1 show the result of the precision and recall obtained from using ranked triple to retrieve result after ranking using Levenshtien reserve engineering approach in comparison with relation ranking approach in FREyA.

Table 1: Evaluation of the proposed triple ranking effectiveness of retrieved res

Characteristic statistics	Precision	Recall
Levenshtien reverse engineering Triple ranking	0.49	0.57
Relation ranking in FREyA	0.45	0.51

Table 1 show that the proposed triple ranking based on levenshtien reverse engineering has outperformed relation ranking in FREyA in terms of the effectiveness of the retrieved result by both the systems. The proposed Levenshtien reverse engineering approach has precision 0.49 and recall of 0.57. While relation ranking has precision 0.45 and recall of 0.51.

CONCLUSION

In this study, we have presented a new semantic triple ranking method based on levenshtien string matching algorithm a reverse engineering approach. The method focuses on semantic ranking of triples generated from natural language. The result shows that ranking triple based on reverse engineering approach has better result than ranking only relation in attempt to use semantically formulated triple for retrieval of relevant result using natural language query.

The result of the retrieve result of the experiment is without resolving ambiguity in the natural language query. Most of the query returned more than possible triple representation of the query due to lack of disambiguation process in the approach. Future research will involve disambiguation of ambiguity in the natural language query before triple ranking.

REFERENCES

Aidan, H., H. Andreas and D. Stefan, 2006. ReConRank: A scalable ranking method for semantic web data with context. Proceeding of the 2nd Workshop on Scalable Semantic Web Knowledge Base Systems.

Anna, F., 2012. Linked data, ontologies and services. Proceeding of the Workshop on Semantic Web Technology, Kuala Lumpur.

Damljanovic, D., M. Agatonovic and H. Cunningham, 2012. FREyA: An interactive way of querying linked data using natural language. In: Garcia-Castro, R. *et al*. (Eds.), ESWC, 2011 Workshops. LNCS 7117, Springer-Verlag, Berlin, Heidelberg, pp: 125-138.

Franz, T., A. Schultz, S. Sizov and S. Staab, 2009. TripleRank: Ranking semantic web data by tensor decomposition. Proceeding of the 8th International Semantic Web Conference (ISWC' 2009), pp: 213-228.

Ivan, H. and K. Miloslav, 2013. SWSNL: Semantic Web Search using Natural Language. Expert Syst. Appl., 40(9): 3649-3664.

Kemafor, A. and S. Amit, 2002. The ρ operator: Discovering and ranking associations on the semantic web. SIGMOD Rec., 31(4): 42-47.

Lixin, H. and C. Guihai, 2006. The HWS hybrid web search. Inform. Software Tech., 48(8): 687-695.

Ramakrishnan, C., W. Milnor, M. Perry and A.P. Sheth, 2005. Discovering informative connection subgraphs in multi-relational graphs. ACM SIGKDD Explor. Newslett., 7(2): 56-63.

Sreeja, G., V.P.E. Eshanna, J. Nidhi and D. Anmesh, 2012. An ontology driven E-counseling system as an implementation of semantic web technology. Proceeding of the International Conference on Computer, Electrical, Electronics and Biomedical Engineering (ICCEEBE'2012). Penang, Malaysia.

Tim, F., D. Li, P. Rong, J. Anupam, K. Pranam, J. Akshay and P. Yun, 2005. Swoogle: Searching for knowledge on the semantic web. Proceeding of the 20th National Conference on Artificial Intelligence. Pittsburgh, Pennsylvania, pp: 1682-1683.

Vanessa, L., P. Michele and M. Enrico, 2005. AquaLoq: An ontology-portable question answering system for the semantic web. Proceeding of the European Semantic Web Conference (ESWC, 2005). Crete, pp: 546-562.

Vanessa, L., F. Miriam, M. Enrico and S. Nico, 2010. PowerAqua: Supporting users in querying and exploring the semantic web. Semantic Web, 3(3).