SEMANTIC AMBIGUOUS QUERY FORMULATION USING STATISTICAL LINGUISTICS TECHNIQUE

Rabiah A.Kadir[1], *Rufai Aliyu Yauri*[2], *Azreen Azman*[3]

[1]Institute of Visual Informatics, Universiti Kebangsaan Malaysia, Malaysia
[2]Department of Computer Science,
Kebbi State University of Science and Technology, Nigeria
[3]Department of Multimedia,
Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia

Email: rabiahivi@ukm.edu.my[1], rufaialeey@yahoo.com[2], azreen@upm.edu.my[3]

DOI: https://doi.org/10.22452/mjcs.sp2018no1.4

ABSTRACT

Natural language query systems mitigate the complexity of structured query. Usually, natural language processing is implemented to solve several problems, such as information retrieval. However, problems such as natural language ambiguity remain unsolved due to the complexity of natural language itself. This issue thus requires further research. Recent studies on semantic query formulation have attempted to resolve ambiguous natural language by proposing different disambiguation approaches. Most such processes are either implemented manually or semi-automated. In the same vein, most recent systems solve ambiguity by using an external dictionary such as WordNet or by providing suggestions manually. The present research proposes a statistical linguistic technique for solving the problem of ambiguity automatically. The proposed technique is experimentally tested on a Quran ontology with queries from the Islamic Research Foundation Website and increases the result of precision and recall by 6% and 10%, respectively.

Keywords: Statistical Linguistics Technique, Semantic Technology, Information Retrieval, Ontology, Islamic Knowledge

1.0 INTRODUCTION

The exponential growth of the number of documents on the web has posed the challenge of determining the relevance of information that can be retrieved from ambiguous query. Search engines were presented as a means of easy retrieval of such documents. The popularity of search engines has revolutionised the way people access and use information on the web in such a way that people today argue that the Internet is Google. However, the popularity of search engines is threatened by the growth of information deposited on the web and the complex information needed by users, which comprises many elaborately interconnected parts that can be comprehended and operated only with specific study or knowledge. Moreover, search engines rely on keyword search and thus cannot cope with problems such as natural language ambiguity and reference reconciliation [1]. Most current retrieval systems, such as those in Google and Yahoo, are based on traditional keyword search, which has the conjunction operator 'AND' between terms and does not focus on the ambiguous query problem [2]. However, keyword search leads to the retrieval of many irrelevant results, which may not be of the user's interest. The current work addresses this issue by applying semantic web technology with the proposed technique in handling ambiguous query. This research is an extension of work originally presented in the 2018 Fourth International Conference on Information Retrieval and Knowledge Management (CAMP'18) [3].

Semantic web technology is introduced to overcome the shortcomings of search engines [4]. However, semantic web represents data in Resource Description Framework (RDF) triple structured format, which requires structured query for information retrieval. During the semantic formulation of natural language query to structured query, ambiguity of natural language becomes an issue. Recently, research on semantic query formulation has attempted to resolve natural language ambiguity by proposing different disambiguation approaches [1,2,3]. Most of these processes are conducted either manually or semi-automatically, as in the case of FREyA [5,6,7,17]. The process engages the user in the disambiguation of ambiguous query and hence may be hectic and time consuming. In some cases, the user may even become excited and select information that may produce query that is unreliable in the semantic representation of the query to structured query. A recent system of disambiguation is based on the use of external dictionaries such as WordNet. The disambiguation process engages the users with the manual suggestion which is mapped the concepts suggestion predicate to the disambiguated concept.

To achieve semantic query disambiguation, this research attempts to represent natural language query into Semantic Web. The data on Semantic Web are represented in RDF format. RDF is a W3C recommended language for representing data on Semantic Web and uses an ontology to represent data into the following triple-form representation:

{Subject, Predicate, Object}.

In this format, Subject and Object represent the concepts that appear on the left and right sides of the triple, respectively. Predicate denotes the explicit relationships that exist between Subject and Object and may be represented by a word, phrase or sentence [8]. To achieve the objective of this research, we perform concept notation using n-grams, which involves the annotation of several ontology concepts and addition of relationships between the concepts. These ontology concepts are stored and form a repository called a knowledge base.

This research describes work on semantic ambiguous query formulation that attempts to semantically formulate natural language query to triple-format representation. The formulated triple is used to generate SPARQL, a standard query language recommended by the W3C for RDF triple format and used to retrieve relevant concepts from a knowledge base using a Jena inference engine with a built-in ontology tool such as Protégé. We propose expansion via statistical linguistics technique to resolve ambiguity in natural language query in the process of transforming to triple form. This method is used against an Islamic knowledge base to retrieve relevant verses from the Holy Quran. The proposed system is tested using a Quran ontology published by Leeds University, United Kingdom. Leeds University's Quran ontology is composed of 300 important noun concepts identified from the Holy Quran and approximately 350 relationships that link the concepts. The collection queries for the experiment are queries of ordinary visitors of the Islamic Research Foundation website, a platform through which people send queries related to Islamic knowledge for answering by expert Dr. Zakir Naik. Then, the experiment is evaluated in terms of capability to retrieve relevant verses from the Holy Quran.

Section 2 discusses related works on semantic query formulation in general and several triple formulation systems. Section 3 explains the implementation of the proposed formulation technique using statistical linguistics. The experiment is analysed and its precision and recall percentage are presented in Section 4. Finally, conclusions are drawn together with recommendations for possible improvement in Section 5, especially the capability of the automated transformation of ambiguous natural language query into structured format.

2.0 SEMANTIC QUERY FORMULATION

Recently, large amounts of information have become available on the web, databases, knowledge bases and related document storage systems. Numerous organisations use automated information systems mostly for information storage. Most important decisions are based on adequate information support, which remains a problem that is based on complex schemata. This issue has led several researchers to focus on query formulation [6]. Natural language query formulation can be conducted semantically via three approaches, i.e. manual, semi-automated and automated.

Manual **-** The system requires the user to manually formulate his query by constructing a particular query language such as SPARQL. Most manual semantic query formulation systems are in visual query formulation (VQF) approach [3]. VQF is a SPARQL query editor that allows the user to manually construct a semantic query. It requires the user to participate in formulating the query by sketching the query [8].

Semi-automated – This approach requires interaction between a computer and the user to semantically formulate natural language query [9]. Two approaches can be implemented; firstly, the system allows the user to select either the concepts or variables before query formulation [10]. Secondly, the user enters a natural language query, and then the system produces variables as suggestions from which the user can choose [11]. Finally, the system uses the selected variables to formulate the query. In this work, ontology is used as a knowledge base structure, where the system uses a high-usability graphical user interface to construct advanced queries. The above system is an improvement from manual semantic query formulation systems in that the user is supported in formulating natural language queries. Instead of having the user engage with complete semantic query formulation, the system allows the user to cooperate during query formulation.

Automated – The system automatically formulates pure natural language query, which is then passed to the system to retrieve the answer from the knowledge base. Several systems have been developed since 2005 for the semantic formulation of users' natural language query to structured query. Stratica, Kosseim, & Desai [12] proposed CINDI in 2005. In this approach, the query is transformed into SQL using a semantic template, which is connected to the conceptual knowledge base from a database schema using WordNet. CINDI was followed by ORAKEL in 2008 [13]. This approach translates factoid questions (e.g. 'what', 'who', 'where' and 'which') using full syntax parsing and a compositional semantics approach. In 2013, NeeluNihalani et. al. [14] worked with a semantic query formulation that is based on a domain-independent natural language interface. The system converts natural language query into SQL by employing a semantic matching technique. The system uses WordNet's lexical database for the semantic matching. The system is composed of two components, namely, a pre-processor and a runtime processor. The pre-processor automatically generates the domain dictionary by reading the database schema. This component uses WordNet to create semantic sets for each table and attribute name in the database. Meanwhile, the runtime processor employs expression mapping, stop word removal and semantic matching techniques to convert the query into structured language SQL. The system is based on a single-fragment query. From 2005 to 2013, several systems have been developed, such as QUICK [15], QACID [16], FREyA [17], QASYO [18], PowerAqua [19], PERSON [20] and ONLI [21].

These systems were developed on the basis of simple or single-fragment queries. By contrast, the present research focuses on ambiguous query formulation using a statistical linguistic technique towards complex fragment queries. The following section discusses the implementation of automated semantic ambiguous query formulation.

3.0 STATISTICAL LINGUISTIC TECHNIQUE FOR DISAMBIGUATION QUERY

This section explains the proposed methodology, which is a statistical linguistic technique of formulating ambiguous natural language query. The formulated natural language query is then matched against the knowledge base to retrieve relevant verses from the Holy Quran.

The statistical linguistics technique comprises natural language processing and a statistical learning approach that uses n-grams to semantically formulate natural language query to structured query. Each natural language query is normalised to identify the concepts between the query and the knowledge base. The query normalisation process involves tokenisation of query, removal of stop words and lemmatisation of the query. The system parses the normalised query to part-of-speech tagger that performs some syntactic analyses of the query and assigns part-of-speech to each word in the query. The tagged tokens are then used by the system to automatically identify concepts and detect a possible predicate for the concepts.

Tokens that are identified as a concept are stored in a container called *Concepts* (e.g. *messenger/NN*). Meanwhile, tokens that are not identified as a concept are stored in another container called *Predicates* (e.g. *have/VBP*). The system uses the list of nouns in *Concept containers* as an input for the system to match between the noun query token with a gazetteer list [22, 23]. The gazetteer list contains 300 noun concepts from the Leeds University Quran ontology.

To handle the ambiguous query, whose concept cannot be identified by the system, this research applies query expansion. It is an effective technique that is mainly used to add new useful words to a user query to improve retrieval [24]. This research implements automated query expansion to expand user queries by adding synonyms of the query tokens during concept identification. The query expansion uses lexical ontologies such as WordNet, semantic parsing of questions based on rules and shared dependency parse trees between the query and the candidate answers [25]. WordNet is an English language-based lexical database for English that symbolises each group of English words into sets of synonyms called synsets, which express the semantic relations between these synonyms and provide short, general definitions for each word. The main idea behind WordNet is actualising the combination of dictionary and thesaurus to support automatic text analysis and artificial intelligence applications. In this experiment, WordNet is used for word disambiguation during concept identification. In this case, the system does not merely perform syntactic matching but also investigates semantic similarities between potential queries and ontology concepts in the gazetteer list. The system automatically attempts to identify *noun* concepts from query tokens by matching any potential concept or its synonyms against the gazetteer list. Fig. I shows the framework followed by the query expansion approach in solving ambiguous query in concept identification.

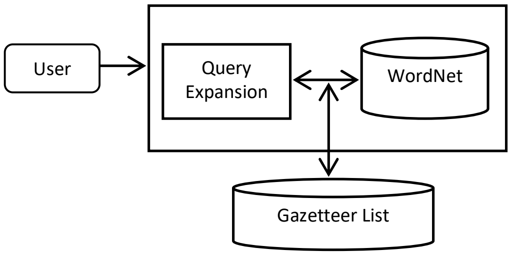

Fig. I: Query Expansion in Concept Identification

Although WordNet contains a sizable number of sufficiently common words, it does not cover for any special domain vocabulary such as the Quran. In this case, equivalent assertion is developed to enrich the query expansion. For example, the query token may express the word *God*; by contrast, in the Quran, Hadith and other Islamic resources, *God* is called *Allah*. The word *Allah* in WordNet is not a synonym of *God*. To deal with this concern, we use the equivalent assertion capability of ontology to indicate that *Allah* is equivalent to *God* in the knowledge base. Equivalent assertion in ontology comprises mechanisms that enable the use and representation of one ontology concept to be equivalent to the other. This process allows the system to process any word or concept as long as they have the same meaning. Therefore, users can input either *Allah* or *God* in his query, and the system recognises both concepts and thus processes them as one. The relation equivalent is shown in Fig. II.

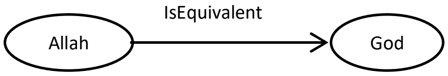

Fig. II: Equivelent Class Assertion

To automatically generate a triple representation (*Subject, Predicate, Object*) in a semantic query formulation, the task of concept identification automatically generates Subject and Object out of query tokens and leaves the remaining task of identifying any possible relationship between the identified concept (predicate). Fig. III shows the relation between subject, object and predicate, which completes triple generation. The system can automatically detect a possible predicate (relation) that depicts the relationship between the subject and object.

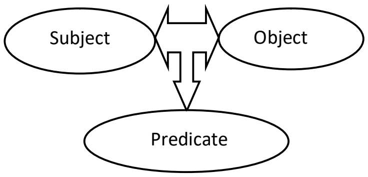

Fig. III: Predicate Detection

The operation of predicate detection involves the use of remaining query tokens to compute predicates using a supervised learning approach with n-gram maximum likelihood estimation. The system uses a training set as an example to do likelihood estimation of possible predicate between identified concepts. The training set contains all possible triple relations for the identified concepts. For example, the system identifies (*Muhammad, Messenger and God*) as concepts, and all possible triple relations are stored in the training set with three difference scenarios, as shown in Fig. IV. In the scenario of triple relations, where the system learns to detect two or more relevant predicates on the basis of user query, the subject and object are identified concepts in the training set.

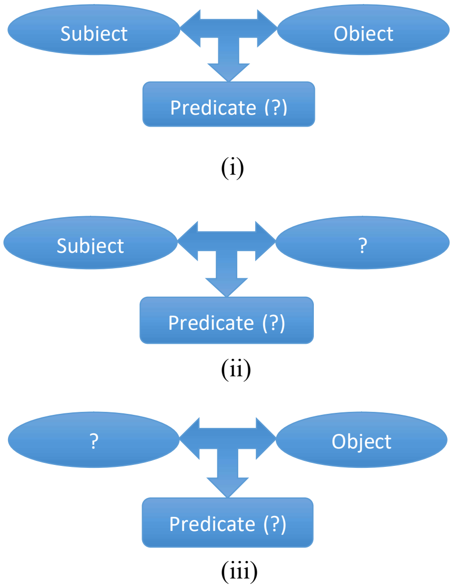

Fig. IV: Triple Relation for Predicate Detection

The system detects the relevant predicate by implementing the n-gram maximum likelihood estimation model towards the triple relations in the training set. The model estimates the probability of the predicate lexicon between the given concepts of computing the probability of a word *w*, from the query with the possible predicate *h* or $P(w|h)$. The system computes all tokens except to identify concepts in the query with a possible predicate in the training set by implementing Eq. (1).

$$P(w_1, w_2, \dots w_n) = \Pi P(w_i) \qquad (1)$$

The system does maximum likelihood estimation by computing the predicate lexicon in the training set and normalising the computation to be within the range of 0–1. The score of the computation is normalised by computing the n-gram probability of a word w_n and that of the previous word w_{n-1}. The system continuously computes the n-gram $c\ (w_{n-1}, \mathrm{w}_n)$ and normalises it by the sum of all the n-grams that share the same first word w_{n-1} as Eq. (2).

$$P(w_n / w_{n-1}) = \frac{C(w_{n-1}\, w_n)}{C(w_{n-1})} \qquad (2)$$

Finally, for this stage, the proposed system produces the relevant triple relations used to retrieve verses semantically from the Holy Quran.

4.0 RESULT AND DISCUSSION

This section explains the evaluation of the proposed statistical linguistics technique in semantic query formulation for verse retrieval of the Holy Quran. The statistical linguistics technique accepts ambiguous and non-ambiguous natural language query and semantically transforms the query into a structured query to facilitate knowledge retrieval from the Holy Quran. For ambiguous query, the query concept is expanded to reduce the number of ambiguous queries in formulating natural language queries using the proposed method. Table 1 shows the sample list of ambiguous queries and its computation using statistical linguistics in giving the expansion of the original query. This expansion is applied in document retrieval by evaluating through precision and recall, as shown in Table III.

Table I: List of Ambiguous Queries

Query	Compute Ambiguous Word
Qur'an says that Allah has made the earth for you as a carpet. This gives an indication that the **earth** is flat. Does this not contradict established modern science?	$P(\text{earth}/\text{world}) = \text{Count}\,\frac{(\text{earth, world})}{\text{world}}$
I have been Wanting to know of other ayat in the Qur'an/Hadiths that are clearly in support of Prophet Muhammad as **last** prophet.	$P(\text{last}/\text{end}) = \text{Count}\,\frac{(\text{last, end})}{\text{end}}$

Table II shows the experimental incapability of the query expansion method to resolve the problem of ambiguous query. The total number of ambiguous queries for this experiment is 28.

Table II: Ambiguous Query Analysis

Methods	No. of Ambiguous Queries	Percentage of Disambiguation Query (%)
Statistical Linguistics Technique + Query Expansion	6	78.6
Feedback + query refinement + vocabulary extended (FREyA)	28	0.0

The evaluation is measured by analysis of the numbers of queries whose ambiguity cannot be solved by the system in the query. The proposed technique fails to automatically resolve six ambiguous queries. This finding shows that the proposed method can solve 78.6% of ambiguous queries, unlike FREyA, whose non-resolution rate is 100%.

The proposed methods in this article enhance the result of solving the identified problems. To perform the experiment of the proposed system, the Quran ontology is developed and applied. The Quran ontology is annotated and stored in Protégée ontology editor, which serves as a knowledge base. A comprehensive evaluation is conducted by comparison of the proposed methods with the technique implemented in FREyA. The comparison is based on the relevance of the returned Holy Quran verses. Then, the effectiveness of the semantically retrieved Holy Quran verses is evaluated using precision and recall.

$$\Pr ecision = \left| \frac{\{relevant\ \ documents\ \} \cap \{retrieved\ \ documents\ \}}{|\{retrieved\ \ documents\ \}|} \right| \tag{3}$$

$$\operatorname{Re} call = \left| \frac{\{relevant\ \ documents\ \} \cap \{retrieved\ \ documents\ \}}{|\{relevant\ \ documents\ \}|} \right| \tag{4}$$

Recall determines how many of the relevant documents are retrieved, as seen in Eq. (3). Meanwhile, Precision calculates how many of the retrieved documents are relevant, as seen in Eq. (4). Measurements of Precision and Recall refer to returned corresponding verses.

Table III: Result of Document Retrieval

Type of Retrieval	Precision	Recall
Statistical Linguistics	0.74	0.87
FREyA Automated Triple	0.68	0.77

Table III shows the performance of the proposed method of natural language query in retrieving relevant verses of the Holy Quran semantically. The experiment is tested on 82 queries, including 28 ambiguous queries. The result shows that Precision and Recall are 0.74 and 0.87, respectively. Meanwhile, FREyA shows Precision of 0.68 and Recall of 0.77. The increments in Precision and Recall are based on the capability of the proposed method to disambiguate the ambiguous query in concept identification.

5.0 CONCLUSION

The main purpose of this research is to resolve ambiguity in queries using query expansion via a statistical linguistic technique during semantic query formulation to retrieve relevant answers from an ontology of the Holy Quran. The University of Leeds' Quran ontology is used as a test bed for this experiment. It provides 300 concepts and 350 relationships, which are mostly being-a, sub-concept relationships.

Semantic formulation of natural language query to structured query involves the transformation of natural language into formal structured query language, which enables the retrieval of semantically structured data in RDF triple format. Structured query language such as SPARQL and SeRQL requires complex syntax, which is challenging to users who are unfamiliar with this structured complex syntax. Simplifying access to semantically structured data requires the system to semantically accept natural language query and transform such query to structured formal query for retrieval purposes. The experiment conducted in this research adopts statistical linguistics to formulate natural language query to structure formal query language. To improve retrieval capability, query expansion is implemented for ambiguous queries in the statistical linguistics technique. The formulated query is used against the knowledge base to retrieve relevant verses of the Holy Quran that will be presented to users. The proposed method shows that the overall result increases by 6% and 10% for Precision and Recall, respectively.

The proposed system in this research does not cover true/false questions, which are popular in Islamic-related queries. In our future work, we intend to explore true/false and negation queries and explore another feature direction, which is the incorporation of Hadith into the semantic Quran search system for the development of an Islamic knowledge base.

ACKNOWLEDGEMENT

This work was supported by the Universiti Kebangsaan Malaysia with grant code FRGS/1/2017/ICT04/UKM/02/8 from the Ministry of Higher Education.

REFERENCES

[1] R. Bentrcia, S. Zidat & F. Marir, "Extracting Semantic Relations from the Quranic Arabic based on Arabic Conjunction Patterns," *Journal of King Saud University – Computer and Information Sciences*, Vol. 30, no. 3, 2018, pp. 382-390.

[2] W. Marlow, "How to Deal with Ambiguous Keyword when Doing Keyword Research," https://willmarlow.com/how-to-deal-with-ambiguous-keywords-keyword-research/ , 2014 (Accessed on 18 August 2018)

[3] R.A.Kadir, A.R. Yauri & A. Azreen, "Automated Semantic Query Formulation for Document Retrieval," in *Proceedings of 2018 Fourth International Conference in Information Retrieval and Knowledge Management* - CAMP'18, 2018, pp. 123-130.

[4] J.M. Valejo, J.G. Ramos Diaz & J.C. Olivres Rojas, "Construction of Semantic Web Search Engine fro a specific Context," in *Proceedings of 2016 IEEE International Autumn Meeting on Power, Electronics and Computing - ROPEC 2016*, 2016, https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7830640

[5] A. Hofstede, H. Proper, & T. Weide, "Computer Supported Query Formulation in an Evolving Context," in *proceedings of Australasian Database Conference*, 1995.

[6] A. Soylu, G. Martin, H. Ian, J. Ernesto, K. Evgeny, Z. Dmitriy, "A Preliminary Approach on Ontology-Based Visual Query Formulation for Big Data," *MTSR*, 2013, pp. 201-212

[7] K. Munir, M. Odeh, P. Bloodsworth, P. & R. McClatchey, "Using assertion capabilities *of an* OWL-based ontology *for* query formulation*,"* in *Proceedings of Information and Communication: From Theory to Applications - ICTTA 2008*, 2008.

[8] P. Ciccarese, M. Ocana, C. Garcia, S. Das, & T. Clark, "An open annotation ontology for science on web 3.0", *Journal of Biomedical Semantics*, Vol. 2 no.2, 2011.

[9] B. Bartosz, B. Marian, & G. Tomasz, "Graphical Query Construction over Scientific Data Sets using Semantic Technologies," 2012, http://www.ci.uchicago.edu/escience2012/pdf/escience2012.

[10] B. Giuseppe, M. Dario, & R. Stefano, "Resolving the query inference problem by optimizing query formulation cost," 1992, pp. 1-30, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.45.4684&rep=rep1&type=pdf

[11] A.O. Enikuomehin, & D.O. Okwufulueze, "An Algorithm for Solving Natural Language Query Execution Problems on Relational Databases," *International Journal of Advanced Computer Science and Applications*, Vol. 3, no. 10, October 2012, pp. 169–175.

[12] N. Stratica, L. Kosseim, & B.C. Desai, "Using semantic templates for a natural language interface to the CINDI virtual library," *Data & Knowledge Engineering*, Vol. *55, No.* 1, 2005, pp. 4–19.

[13] P. Cimiano, P. Haase, J. Heizmann, M. Mantel, & R. Studer, "Towards portable natural language interfaces to knowledge bases – The case of the ORAKEL system," *Data & Knowledge Engineering*, Vol. 65, No. 2, 2008, pp. 325–354

[14] N. Neelu, S. Sanjay, & M. Mahesh, "Natural language Interface for Database: A Brief review," *International Journal of Computer Science Issues*, Vol. 8, No. 2, 2011, pp. 600-608

[15] G. Zenz, X. Zhou, X., Minack, E., Siberski, & W. Nejdl, "From keywords to semantic queries—Incremental query construction on the semantic web," *Web Semantics: Science, Services and Agents on the World Wide Web*, Vol. 7, No. 3, 2009, pp. 166–176

[16] Ó. Ferrández, R. Izquierdo, S. Ferrández, & J.L. Vicedo, "Addressing ontology-based question answering with collections of user queries," *Information Processing & Management*, Vol. 45, No. 2, 2009, pp. 175–188

[17] D. Damljanovic, M. Agatonovic, & H. Cunningham, "Natural Language Interfaces to Ontologies : Combining Syntactic Analysis and Ontology-Based Lookup through the User Interaction," in *Proceedings of the Semantic Web Research and Applications*, 2010, pp. 106–120.

[18] A.M. Moussa, & R.F. Abdel-kader, "QASYO : A Question Answering System for YAGO Ontology," *International Journal of Database Theory and Application,* Vol. 4, No. 2, 2011, pp. 99–112.

[19] V. Lopez, M. Fernández, N. Stieler, E. Motta, W. Hall, M.K. Mkaa, & U. Kingdom, "PowerAqua : supporting users in querying and exploring the Semantic Web," *Semantic Web Journal*, 2011.

[20] A. Aksac, O. Ozturk, & E. Dogdu, "A novel semantic web browser for user centric information retrieval: PERSON," *Expert Systems with Applications*, Vol. 39, No. 15, 2012, pp. 12001–12013

[21] C. Unger, L. Bühmann, J. Lehmann, A.C. Ngonga Ngomo, D. Gerber, & P. Cimiano, "Template-based question answering over RDF data," in *Proceedings of the 21st international conference on World Wide Web - WWW '12*, 2012, pp. pp. 639-648

[22] N. Suryana, F.S. Utomo & M.S. Azmi, "Quran Ontology: Review on Recent Development and Open Research Issues," *Journal of Theoretical and Applied Information Technology*, 2018, Vol. 96 no. 3, pp. 568-581.

[23] S. Dlugolinsky, . Nguyen, M. Laclavik, & M. Selang, "Character Gazetteer for Named Entity Recognition with Linear Matching Complexity," in *Proceedings of 2013 3rd World Congress on Information and Communication Technology - WICT 2013,* December 2013

[24] M. Mandar, S. Amit, & B. Chris, "Improving Automatic Query Expansion," in *proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '98*, 1998, pp. 206-214.

[25] C. Carpineto, & G. Romano, "A Survey of Automatic Query Expansion in Information Retrieval," *ACM Computing Surveys*, Vol. 44, No. 1, 2012, pp. 1–50