UNIVERSITI PUTRA MALAYSIA

AUTOMATED SEMANTIC QUERY FORMULATION FOR QURANIC VERSE TRANSLATION RETRIEVAL

ALIYU RUFAI YAURI

FSKTM 2014 8

AUTOMATED SEMANTIC QUERY FORMULATION FOR QURANIC VERSE TRANSLATION RETRIEVAL

By

ALIYU RUFAI YAURI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirement for the Degree of Doctor of Philosophy

August 2014

Abstract of this thesis presented to the Senate of the University of Putra Malaysia in fulfilment of the requirements for the degree of Doctor of Philosophy

AUTOMATED SEMANTIC QUERY FORMULATION FOR QURANIC VERSE TRANSLATION RETRIEVAL

By

ALIYU RUFAI YAURI

August 2014

Chairperson: **Rabiah Abdul Kadir, PhD**

Faculty: **Computer Science and Information Technology**

With the exponential growth in the amount of data that is deposited on the web and in other data storage repositories daily, there is an increase in the global desire to retrieve that data in a more effective and efficient manner. There are quite a number of mechanisms through which this data is retrieved, such as search engines like Yahoo and Google among others, however most of the current information retrieval mechanisms on the web are based on a keyword search. A keyword search mostly retrieves information that is not relevant to the searched query due to problems such as semantic ambiguity of natural language. The user needs to know the exact keyword to use in order to retrieve the relevant information. To overcome this problem, several approaches have been researched, such as the query formulation, and most are based on a keyword and small fragment query.

In this thesis, a study of the automatic semantic query formulation of natural language query to structured query is proposed. The proposed system in this thesis is referred to as AutoSQuR, meaning *Automated Semantic Quran Retrieval*. The proposed AutoSQuR attempts to semantically formulate complex natural language queries to triple representation and retrieve relevant verses from Holy Quran.

The main contribution of this research is introduced a method to formulate semantic query automatically for natural language queries to structured queries using statistical machine learning technique. The contribution includes going beyond keywords and formulating small fragment queries to complex queries that can be a paragraph in length. Additionally the proposed system supports both categories of users who prefer suggestions from the system and those who prefer to reformulate their query in case the system fails to automatically formulate user queries. The proposed system provides suggestions to the user where either concepts are identified or not in the query. Another contribution is the use of ontology equivalent assertions due to the limitations of WordNet for the disambiguation of Islamic-related words.

Finally, an experimental evaluation of AutoSQuR is implemented. The evaluation was based on measuring the performance of the proposed statistical machine learning technique with the existing approach in FREyA in terms of the percentage of queries that are semantically formulated correctly, and the effectiveness of the retrieved Quran verses. Evaluation has shown that the proposed approach outperformed the existing approach in FREyA. The statistical machine learning technique has shown improvement of 17.4% increases in comparison with the existing approach in FREyA in terms of correctness of the query formulation. Meanwhile, in the effectiveness of the retrieved verse, the proposed approach shows an improvement of 0.06 in terms of precision and 0.1 for recall.

`

Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Doktor Falsafah

AUTOMATIK FORMULASI QUERY SEMANTIK UNTUK AL AYAT TERJEMAHAN PENCARIAN

Oleh

ALIYU RUFAI YAURI

Ogos 2014

Pengerusi: **Rabiah Abdul Kadir, PhD**

Fakulti: **Sains Komputer dan Teknologi Maklumat**

Dengan pertumbuhan yang pesat pada jumlah data yang terdapat di web dan repositori simpanan data yang lain setiap hari, terdapat peningkatan dalam keinginan global untuk mengambil data tersebut mengikut cara yang lebih berkesan dan cekap. Terdapat beberapa mekanisma enjin carian seperti *Yahoo, Google* dan lain-lain di mana data tersebut boleh dicapai, walau bagaimana pun kebanyakan daripada mekanisma dapatan semula maklumat pada web adalah berdasarkan pada carian kata kunci. Carian kata kunci kebanyakannya dapat semula maklumat yang tidak berkaitan dengan pertanyaan pencarian kerana masalah seperti semantik dan kekaburan bahasa tabii. Pengguna perlu mengetahui kata kunci yang tepat untuk digunakan dalam usaha mendapatkan semula maklumat yang berkaitan. Untuk menangani masalah ini, terdapat beberapa penyelidikan telah dilakukan seperti pengungkapan pertanyaan yang kebanyakannya berasaskan kata kunci dan cebisan pertanyaan yang mudah.

Dalam tesis ini, satu kajian pengunkapan semantik bagi pertanyaan bahasa tabii kepada pertanyaan berstruktur dicadangkan secara automatik. Sistem yang dicadangkan dalam tesis ini dirujuk sebagai AutoSQuR, yang bermaksud *Automated Semantic Quran Retrieval*. AutoSQuR yang dicadangkan cuba untuk mengunkap pertanyaan bahasa tabii yang kompleks kepada perwakilan *triple* secara semantik dan mendapat semula ayat-ayat yang berkaitan daripada kitab suci Al-Quran.

Sumbangan utama kajian ini adalah memperkenalkan satu kaedah untuk mengunkap semantik pertanyaan secara automatik untuk pertanyaan bahasa tabii yang kompleks kepada pertanyaan berstruktur dengan menggunakan teknik pembelajaran mesin statistik. Sumbangan kajian ini merangkumi pemadanan yang melangkaui kata kunci dan pengungkapan cebisan pertanyaan yang mudah kepada pertanyaan kompleks yang panjangnya sehingga satu peranggan. Disamping itu sistem yang dicadangkan menyokong kedua-dua kategori pengguna iaitu memilih cadangan daripada sistem dan memilih untuk mengungkap semula pertanyaan mereka bagi kes sistem yang gagal untuk mengunkap pertanyaan pengguna secara automatik. Sistem yang dicadangkan menyediakan beberapa cadangan kepada pengguna sama ada konsep yang telah dikenal pasti atau sebaliknya. Di antara sumbangan yang lain ialah

penggunaan kenyataan yang setara bagi ontologi kerana kekangan *WordNet* untuk penyahkaburan perkataan yang berkaitan dengan Islam.

Akhir sekali, penilaian eksklusif eksperimen ke atas AutoSQuR dilaksanakan. Penilaian ini adalah berdasarkan pengukuran prestasi teknik pembelajaran mesin statistik yang dicadangkan dengan pendekatan yang sedia ada pada FREyA dari segi peratusan ketepatan pengungkapan pertanyaan secara semantik dan keberkesanan dapatan semula ayat-ayat Al-Quran. Penilaian menunjukkan pendekatan yang dicadangkan mengatasi pendekatan yang wujud dalam FREyA. Teknik pembelajaran mesin statistik membuktikan 17.4% peningkatan dalam ketepatan pengunkapan pertanyaan berbanding dengan pendekatan dalam FREyA. Manakala dalam keberkesanan dapatan semula ayat-ayat, pendekatan yang dicadangkan menunjukkan peningkatan 0.06 dalam kejituan dan 0.1 untuk perolehan.

ACKNOWLEDGEMENTS

First of all, all praise to Allah the most merciful who have given the opportunity to attain to this level.

I wish to express my sincerest gratitude to my supervisors, Dr Rabiah Abdul Kadir, Dr Azreen Azman and Professor Masrah Azrifah Azmi-Murad, who have supported me throughout my studies as warm and supportive guardians. I must acknowledge their support and encouragement which has contributed immensely towards the completion of my PhD.

I cannot end without thanking my family, in particular my father who singlehandedly supported my studies financially and also spiritually. I would also like to thank my entire family and friends for their support throughout my studies.

I certify that a Thesis Examination Committee has met on 29 August, 2014 to conduct the final examination of Aliyu Rufai Yauri on his thesis entitled "Automated Semantic Query Formulation for Quran Verse Translation Retrieval" in accordance with the Universities and University Colleges Act 1971 and the constitution of the Universiti Putra Malaysia [P.U. (A) 106] 15 March 1998. The committee recommends that the student be awarded the Doctor of Philosophy degree.

Members of the Thesis Examination Committee were as follows:

Norwati Mustapha, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Ali Mamat, PhD
Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

Muhamad Taufik Abdullah, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

_____

Noritah Omar, PhD.
Associate Professor and Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:

This thesis was submitted to the senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirements for the degree of Doctor of Philosophy. The members of the supervisory committee were as follows:

Rabiah Abdul Kadir, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Azreen Azman, PhD
Senior Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

Masrah Azrifah Azmi-Murad, PhD
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Member)

BUJANG BIN KIM HUAT, PhD
Professor and Dean
School of Graduate Studies
Universiti Putra Malaysia

Date

Declaration by Graduate Student

I hereby confirm that:

- this thesis is my original work;
- quotations, illustration and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institution;
- intellectual property from the thesis and copyright of the thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- writing permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of writings, printed or in electronic form) including books, journal, modules, proceeding, popular writings, seminar papers, manuscripts, posters, reports, lecture notes, learning modules or any other materials as stated in the Universiti Putra Malaysia (Research Rules 2012;
- there is no plagiarism or data falsification/fabrication in the thesis, and scholarly integrity is upheld as according to the Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) and the Universiti Putra Malaysia (Research) Rules 2012. The thesis has undergone plagiarism detection software.

Signature: _____ Date: _____

Name and Matric No: Aliyu Rufai Yauri (GS27702)

Declaration by Members of Supervisory Committee

This is to confirm that:

- The research conducted and the writing of this thesis was under our supervision;
- Supervision responsibilities as stated in the University Putra Malaysia (Graduate Studies) Rules 2003 (Revised 2012-2013) are adhered.

Signature: _____
Name of
Chairman of
Supervisory
Committee: _____

Signature: _____
Name of
Member of
Supervisory
Committee: _____

Signature: _____
Name of
Member of
Supervisory
Committee: _____

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

1.1 Background of Study

Users today rush to various available search engines to search for information whenever the need arises. Although these search engines are able to return varied information related to search queries, many returned results are not relevant. In most cases users need to navigate through several pages before they can get what they need. Sometimes a query cannot be answered by a single document, instead several documents needs to be navigated before needs are fulfilled. This may be because the required information involves multiple sub-topics, and documents relevant to different sub-topics are possibly diversified in the returned documents (Dongyi Guan,2013). In the end, a user may end up not fully satisfied with what is presented to them. This is because a computer doesn't understand the meaning of the query and thus, goes for traditional keyword searching. In a traditional keyword search the user needs to know the exact keywords to use in order to retrieve relevant information, and computer perform keyword matching without having the knowledge of what the query means. For example, a user may be trying to search for the cost of a Jaguar car and pose their query using the words "What is the cost of jaguar?" .In the traditional keyword search, the search engine performs keyword matching of the query words against the document and retrieve contents that are associated with the set of keywords. Here search engines may return information about jaguar cars, jaguar animals, or products named *jaguar*, among other things, without considering the true meaning of the user, which is *Jaguar car*, because computers don't know the difference between a jaguar as an animal or as a car, and thus the user is left with many documents to go through in order to get to the desired information.

To overcome the shortcomings of the traditional keyword search system, the concept of the semantic web was introduced by the W3C consortium. The semantic web, in other words a web of linked data, is an extension of the current version of the web whereby information is given a well-defined meaning to enable human and computers to easily work together. Semantic Web Technology proposed solutions to the current limitations of web search systems by extending the current web. In the semantic web approach data is given a well-defined format that models the meaning of information on the web, as well as applications and services, so as to discover, annotate, process and publish data that is encoded in them (Zou, Finin, & Chen, 2004.). This is done to facilitate semantic searches of well-defined data, where computers understand the user's query intention and retrieve corresponding results based on matching concepts rather than keywords (Solskinnsbakk, 2012). In semantic web user query is represented by the same format as the document for the computer to easily understand and process. The user query is transformed into data format before being matched against a corresponding document for retrieval. This will enable the retrieval of results that are more relevant to the user compared with the traditional keyword search. The concept of the semantic web is involved in many

areas such as knowledge representation and reasoning, information retrieval, databases, natural language processing, and machine learning among others.
The data in the semantic web is represented into Resource Description Framework (RDF) format. RDF is a W3C recommended language for representing data on the semantic web. RDF uses ontology to transform data into graphical triple form representation.

{Subject, Predicate, Object}

In simple terms ontology can be seen as objects that may exist in a particular domain and the relationships that may exist between the objects. In practice, objects and concepts are used to represent the same thing when dealing with ontology. We will be using concepts throughout this thesis. In RDF concepts are represented in triple form, which shows how two concepts are related to each other. In triple format, a subject represents a concept that appears on the left side of the triple, an object is a concept that appears on the right hand side of the triple, and a predicate stands for explicit relationships that exist between subject and object which may be represented by a word, phrase or sentence. In this thesis we will be interchanging predicates and relationships to describe relationships between concepts. Data represented in RDF triple format is stored in the knowledgebase in order to facilitate manipulation and querying. When a particular domain is identified, all concepts in that domain are identified and annotated, and this is then stored in the knowledgebase. Concept notation is the process of performing several types of annotation on ontology concept such as adding relationships between concept comments (Ciccarese, Ocana, Garcia, Sudeshna & Clerk , 2011). When these ontology concepts are annotated and stored, the repository in which they are stored is referred to as the knowledgebase.

The challenge is that querying data represented in RDF structure in the knowledgebase requires a structured query such as SPARQL. SPARQL is the standardised query language recommended by W3C for RDF triple format, which is used for manipulation and retrieval in the same way SQL is the standardised query language for relational databases. User's natural language queries need to be semantically formulated to the same structure as the information stored in the knowledgebase. Semantically annotated ontology stored in the knowledgebase relies on a semantically formulated query in order to retrieve the relevant information. Computers can retrieve more relevant documents by focusing on the underlying meanings in queries. The structured representation of the query enables computers to understand true intentions more precisely, and effectively retrieves more relevant documents from the knowledgebase. These structured queries are represented in complex syntax which requires the user to be familiar with how to use the syntax before they can use the query language for retrieving the desired information from the knowledgebase. Recent studies show that users prefer using natural language to structured query language (Tablan, Damljanovic & Bontcheva, n.d., Kaufmann & Bernstein, 2008). Natural language query systems hide the complexity of the structured query thereby enabling users to use natural language in order to retrieve information by semantically formulating the query into a structured query. However, due to the complexity of natural language where problems such as natural language

ambiguity remain unsolved, the area remains subject to research. Search engines such as Google have incorporated the semantic web into their search processes, as with the knowledge graph introduced by Google, where they have started translating their data into triple form representation. The key idea of knowledge graphs is to enable Google's search process to have a better understanding of the user query, so that relevant content around the main topic is more likely to be presented to the user. The process of processing queries is still based on a traditional keyword search (Amit Singhal, 2012).

This thesis describes work on the automatic semantic query formulation of a Natural language query that attempts to semantically formulate a natural language query to triple representation of the query. The formulated triple is then used to generate a SPARQL query which is matched against the knowledge base to retrieve relevant results. The proposed system is referred to in this thesis as "AutoSQuR", meaning Automated Semantic Quran Retrieval. AutoSQuR was implemented using the semantically formulated query based statistical machine learning approach to query and exploring Quran ontology. The research shows how a natural language query is automatically transformed to a structured triple representation of the query, which is used against the Quran knowledgebase in order to retrieve relevant verses from the Holy Quran. The approach involves the use of Quran ontology, which is used as the knowledge base for the system. A semantic query formulation module adopted statistical machine learning techniques for automatic formulation of natural language query to structured triple form representation. The formulated triple is used to generate SPARQL to retrieve the Quran knowledgebase using the Jena inference engine.

The experiment to be undertaken will not focus on both complex and simple queries, it will not tackle queries that return Boolean True/False answers. Our main focus is to present verses retrieved from the Quran that answer semantically formulated queries. In Boolean queries the main target is to answer a user with either true or false, which is not the focus of our research. The proposed system will not elicit a yes/no answer.

The system was tested using the Quran ontology published by Leeds University of the 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 (Dukes & Atwell, 2009). The query set used for the experiment in this research was obtained from the Islamic Research Foundation Website where people send their queries related to Islam and experts answer the questions. Expert Quran judgment on the queries was given by Dr Kabiru Goje from Universiti Sains Islam Malaysia, and was used as a benchmark to evaluate the proposed AutoSQuR approach in this research. The proposed AutoSQuR was also evaluated by comparing the answers returned with the answers returned by the existing semantic query formulation approach of FREyA (Damljanovic, Agatonovic, and Cunningham, 2012) and the traditional keyword-based Quran retrieval system, Qurany(Abbas and Atwell, 2012). Although Qurany is a traditional keyword search system which main motive is to match query keywords against Quran in order to

retrieve relevant Quran verse not semantic search, we chooses to compare the result of Qurany with that of approach in this research in order to clearly show the limitation of current keyword Web search systems such as goggle, yahoo in terms of retrieving irrelevant result.

1.2 Motivation

Getting computers to understand the meaning of a query and corresponding documents was the motivation for the semantic web. The semantic web is getting more popular by the day, due to several research projects that have been undertaken regarding the semantic web. The fact that the semantic web search approach requires the use of structured syntax has limited the utility of the semantic web technology concept. This is because users don't want to go through these complex syntax before they are able to retrieve information for their needs. Most users prefer to use natural language to pose their query rather than structured queries with complex syntax that must be learned (Wang,Yang, Chenglei, , Rui,Meng,Lui, 2007). Although there are existing semantic query formulation systems that semantically formulate natural language queries to structured queries, most of the existing approaches are based on keywords and small fragment queries. Due to the complex nature of natural language, and the desire to get an effective system where users can easily make natural language queries using vocabularies of their choice and get more precise results, the area is still subject to research. In view of this, providing an effective means through which they can use natural language using vocabularies of their choice and retrieve effective results is the main motivation of this thesis.

Another motivation of this thesis is the rise in global interest in Islamic related knowledge, especially from the Quran. We intend to provide an effective tool with an approach that support users with Islamic knowledge or not be able to get their queries answered effectively. In this case both Muslims and non-Muslims without knowledge of the structure of a Quran document can get their natural language query answered.

Users whose queries are related to Islam tend to pose their questions using many words. Users may begin by stating their problems and then following the query with many more words, or make statements and ask their main question at the end. With the current system, such types of query will struggle and as a result answers may not be returned. The main motivation for the work in this thesis is to provide effective approach that accepts natural language queries of any form, where users can use phrases, sentences or paragraphs and retrieve the relevant information in respect to the query. A user's natural language query is semantically formulated to a structured representation such as triple base representation in order to access semantically structured knowledge stored in the knowledgebase. This will facilitate the retrieval of relevant information in an effective and efficient manner.

1.3 Research Problems

The exponential growth of documents on the web has posed the challenge of how best this data could be retrieved. Search engines were presented as a means of easy retrieval of such documents. The popularity of the search engine has revolutionized the way people access and use information on the web in such a way that today people are arguing that the internet is Google. However, the popularity of search engines is threatened with the growth of information deposited on the web, coupled with the complex information needs of the users. Search engines rely on keyword searches and as a result cannot cope with problems such as natural language ambiguity and reference reconciliation. Most of the current retrieval systems such as those in Google and Yahoo, among others, are based on traditional keyword searches, and keyword searches lead to the retrieval of many irrelevant results, which may not be in the user's interest.

Semantic web technology was proposed to overcome the shortcomings of the current keyword based search engines. The semantic web transforms data into a RDF triple structured format that requires a structured query in order to retrieve the desired information, and therefore requires the user to use a complex structured query to retrieve results. The main problem is that before formulating a query using a structured query such as SPARQL, a user needs to know the structural representation of the document, in other words the schema of the document in the knowledgebase, and the syntax of the structured query language (Jarrar, & Society & Dikaiakos, 2012). To overcome the above problem where users need to know the standard structure of the document before they can retrieve the desired information, the system needs to formulate the natural language query into structured triple form. A user's natural language query needs to be semantically formulated in order to be matched with semantically stored information in the knowledgebase for retrieval.

Although there are existing systems that facilitate retrieval from the knowledgebase using natural language queries, such as FREyA, ORAKEL, and QUESTIO among others, they are mostly designed to handle small fragment queries such as phrases and single sentences (Habernal & Konopík, 2013). Most of the current semantic query formulation approaches are based on manual or semi-automatic approach where is heavily involved in the query formulation processing which is hectic and time consuming. Where user is involved in the process of semantic query formulation such selecting concepts from a form based system, and mapping the concepts with relationship. A detailed explanation of semantic query approaches will be discussed in Chapter 2.
Most of the existing approaches struggle with complex natural language queries, i.e. multiple sentences query. When multiple sentence queries are used for those systems, the result is usually not good.

During the process of semantically translating natural language queries to structured queries, the ambiguity of natural language remains an issue. Although recent research into semantic query formulation attempts to resolve natural language ambiguity by proposing different disambiguation approaches, most of

disambiguation processes are either done manually or semi-automatically, as in the case of FREyA. The process engaging the user in disambiguation of ambiguous query, where user is required to manually disambiguate the ambiguous words which is hectic and time consuming. In some cases, users may even get excited and choose information that may end up forming a query that is not the best semantic. In the same vein, most recent system disambiguation is based on using external dictionary such as WordNet, but when dealing with Islamic-related vocabularies WordNet possesses some limitations, for example not all Islamic-related vocabularies are included in WordNet.

Translating a natural language to structured SPARQL query language requires the transformation of the natural language query to triple based format. SPARQL query syntax consists of the set of triples, *subject, predicate, object* where the subject, predicate and/or object can consist of variables. Therefore, the main idea of the semantic query formulation system is to transform a user's natural language query to triple format representation where the variables are then parsed to SPARQL generation. However, the system may attempt to transform natural language to triple based representation but, may end up with more than one possible triple, which thus poses the challenge of identifying the best possible triple based representation of the query. Thus in the case of manual or semi-automatic semantic query formulation, users may be left with many triples to choose from in order to allow for further processing. In terms of automatic approaches the task is even more difficult, since the computer is left with so many triples to process in order to return an answer.

Recent approaches in semantic Query formulation attempt improvements from manual and semi-automatic semantic query formulation methods such as those in AutoSPARQL and DENNA, but in cases where the system fails to automatically formulate a user's natural language query, the system fails without any further processing. Some approaches, such as those in FREyA, engage the user by providing a suggestion when the system fails to automatically formulate a user's natural language query. In FREyA, when the system faces ambiguity in the query, it will not be able automatically semantically formulate the query and thus the user is engaged for disambiguation process. A suggestion is then provided to the user in order to disambiguate the query by manually mapping the concepts in the query with suggested predicate. These suggestions may fail to impress the user and as result the user may try to reformulate the query again and again. Although some users may be happy to get suggestions from the system when the system fails to automatically reformulate their original query, other users may prefer reformulating it themselves instead of being provided with suggestion.

The suggestions provided by the existing system when the system fails to semantically formulate a user's natural language query to a structured format are based on the identification of ontology concepts in the query tokens. In cases where the user's query tokens do not contain any ontology concepts, the current approaches are not able to assist users with suggestions, as in FREyA. When the system is not able to semantically formulate a user's query and fails to provide any suggestions to the user, they may end up reformulating the query several times or even quitting the

search process, assuming such information does not exist in the knowledgebase. It is likely that such information as they are looking for actually exists in the knowledgebase, it is just that the system is built around identifying concepts in the query and then attempting to relate the identified concepts. When concepts are not identified from the query tokens, the existing system will fail to automatically formulate the query, and not provide any form of suggestion to the user. Instead the system just fails and requires the user to either exit or reformulate the query.
Finally, most of the current Quran search are based on traditional keyword matching or concept searches such as in Qurany, which is a popular Quran search system for keywords, based search for Quran verse and concept search. These search approaches lack semantics and thus retrieve a lot of irrelevant information.

1.4 Research Objectives

In order to overcome the research problem outlined in the previous section, this research focuses on several objectives:

(1) To propose an algorithm for automated formulation of natural language query semantically.
(2) To enhance the performance of document retrieval semantically by introducing query disambiguation method and suggestion based approach.
(3) To propose a method to rank the most relevant triple representation semantically.
(4) To develop a prototype that is able to semantically formulate a natural language query that may be of paragraph length, i.e. a query represented as either a phrase, single sentence or multiple sentence query.
(5) To evaluate the prototype through the accuracy of the formulated queries by looking at the precision and recall.

1.5 Research Scope

The proposed approach accepts natural language queries of any length and automatically transforms the query into triple based representation and uses the triple representation to generate a SPARQL query language which is then used against the knowledgebase for retrieval of the answer. The proposed system is designed to accept phrase base queries, single sentence queries and multiple sentence queries. The knowledgebase is made up of noun concepts from the holy Quran. In view of this, the queries are mainly related to noun concepts from the Holy Quran domain. The system does not cover verb-based Islamic-related queries such as those about zinnia or solat among others, however, they could be used as predicates to the concepts.

The system is designed to retrieve relevant answers from the user query by retrieving the most appropriate concept and retrieving corresponding verses for that concept. Boolean queries that answer the user with true/false or elicit yes/no answers are not supported in this research.

1.6 Research Contribution

The main contributions of this research can be described as follows:

(1) The proposed approach goes beyond semi-automatic semantic query formulation to fully automated semantic query formulation. Automating the semantic query formulation process makes it easier to use a natural language query to query structured data without involving users in a complex and time consuming process. The implementation of the automated semantic query formulation was based on statistical machine learning technique using N-gram maximum likelihood estimation. In the proposed method the natural language query is automatically semantically formulated to triple representation of the query, and the triple variables are parsed for SPARQL query generation, which is then matched against the knowledge base, using the Jena Inference engine for retrieval. Details of this contribution will be described in Chapter 3 and Chapter 4.

(2) The proposed approach has the ability to semantically formulate a user's natural language query, retrieve the answer for the semantically formulated query. The answers are represented to the user inform Quran verses related to the given user's natural language query. Semantically formulating the natural language query to a structured query enables translation of the unstructured natural language query to the same structure as the document which allows for retrieval of a document semantically and thus contributes to an increase in the percentage of relevant documents that are retrieved.

(3) The proposed system enables automatic disambiguation of Islamic-related natural language queries using the WordNet Lexical dictionary and ontology equivalent assertions to assert that a particular concept or relationship is equivalent to another concept or relationship. The system automatically disambiguates natural language query tokens based on synonym detection using the WordNet lexical dictionary in the case of ambiguity during concept identifications. When there is ambiguity in the process of uni-gram or bi-gram estimation to detect predicate, the system automatically uses Wordnet to disambiguate ambiguous words. For Islamic-related words that are lacking in WordNet, an ontology-equivalent assertion is used to populate the knowledgebase with equivalent concepts or properties. Equivalent assertion enables retrieval of equivalents of the user given concept, for example, which allow retrieval of equivalents of such concepts in the knowledgebase if the concept does not exist.

(4) In the proposed approach, in cases where the system fails to automatically formulate natural language query to structured query, it doesn't just fail and exit. In the proposed method when a user's natural language query is not automatically formulated, the proposed approach provides a choice for the user to reformulating the query or get suggestions from the system. This allows the user to choose the preferred option and will thus save the user time. For example, when only reformulation option is provided to the user, user may be required to reformulate the query several times if he is not able provide in his query term that can be used by the system to formulate the query. The suggestion proposed in this thesis is based on either concepts being identified in the natural language

query which is different from current approaches which relies on the identification of concepts from the query before any form of suggestion can be presented to the user.

(5) During the process of semantically formulating natural language to structured triple representation, more than one possible triple representation of the query may emerge which will make it difficult to automatically retrieve relevant documents. There is therefore a need to get one most appropriate triple representation of the query in order to automatically retrieve a relevant answer by ranking the returned triples. In a case where the proposed semantic query formulation approach returns more than one possible triple representation of the given natural language query, the proposed approach is semantic triple ranking of the triples based on Levenshtien String matching, a reverse engineering approach. The ranking approach enables ranking of the returned triple representation of the given natural language query in order to get the most appropriate triple representation of the query which allows further processing of the query for automatic retrieval of relevant documents.

(6) The proposed approach has reduced the onerous task of involving the user in the query formulation process by automating the triple format transformation of the user query. The system uses examples from the training set automatically presenting the triple based representation of the user query at once, without taking the concepts one by one, and doesn't require the user to do the choosing of the corresponding relationships.

(7) Improve the precision and recall of the retrieval of Quran verses for natural language queries. The proposed approach in this thesis has shown an overall increase of 6% in term of precision and a 10% increase in terms of recall verses.

1.7 Thesis Organisation

This thesis is organised as follows.

Chapter 2 is an overview of the semantic web. The concept of a semantic search is discussed. An overview of the current trend in semantic query formulation approaches shows that systems today support users posing their queries using natural language queries. A comprehensive review of semantics embedded in information retrieval is made, and a review of the computational tools in the Quran is provided. A summary of the chapter is presented.

Chapter 3 provides the methodology approach used for this research. It shows the statistical machine learning using N-gram maximum likelihood estimation technique used, which is semantically formulate user queries and assist users with options in the event the system fails to automatically formulate a query. An approach to handling situations when the system fails to semantically formulate a natural language query to a structured query is proposed. A new triple ranking approach is also presented in this chapter.

In Chapter 4 a comprehensive detailed step-by-step implementation of how the proposed method of semantic query formulation is implemented in information retrieval is presented. Implementation of the proposed suggestion-based approach is also presented in this chapter. Triple ranking implementation is also presented in this chapter. Additionally, implementation of how the Quran verses related to semantically formulated query is retrieved is presented.

In Chapter 5 comprehensive experiments and the results of the research are presented. In this chapter, evaluation of the semantic query formulation approach, i.e. statistical machine learning is presented in comparison to the existing research on semantic query formulation in FREyA and Traditional keyword based Quran search systems based on using human expert judgments as bench mark.

In Chapter 6 a comprehensive analysis of the experiments and results presented in Chapter 5 is presented. The chapter analyses and discusses the results obtained from the research experiments when compared with the existing approach in FREyA.

Chapter 7 presents the conclusion of the proposed semantic query formulation approach, AutoSQuR, presented in this thesis. The achievements of the research are presented, and some challenges in, and future work on, the research are highlighted.

REFERENCES

A. Doan, J. Madhavan, P. Domingos, A. Halevy. (2004) .Ontology matching: a machine learning approach, in: S. Staab, R. Studer (Eds.), Handbook on Ontologies, Springer-Verlag, Berlin.

A. M. Moussa and R. F. Abdel-Kader. (2011).QASYO: A Question Answering System for YAGO Ontology. International Journal of Database Theory and Application Vol. 4, No. 2, June, 2011

A.H.M. ter Hofstede, H.A. Proper and Th.P. van der Weide (1995) .Computer supported query formulation in an evolving context. In Proceeding Sixth Australasian Database Conference, ADC'95, Volume 17(2) of Australian Computer Science Communications, Adelaide, Australia (January 1995)

Abbas, N., Atwell, E. (2012). Qurany: how to search for concepts rather than words in a corpus. Proceeding IVACS'2012, Leeds, UK.

Abbas, Noorhan Hassan. (2009). Quran 'Search for a Concept' Tool and Website. Msc Thesis. The University of Leeds. UK.

Ahmed, Z., & Gerhard, D. (2007). Web to Semantic Web & Role of Ontology. In the proceedings of National Conference on Information and Communication Technologies, (NCICT-2007), Pakistan, 9th May 2007.

Aksac, A., Ozturk, O., & Dogdu, E. (2012). A novel semantic web browser for user centric information retrieval: PERSON. Expert Systems with Applications, 39(15), 12001–12013.

Al-Gharaibeh, A., Al-Taani, A., & Alsmadi, I. (2011). The Usage of Formal Methods in Quran Search System. ICICS 2011. Retrieved from http://www.icics.info/icics/Proceeding/ICICS.paper/61.pdf

Al-Harbi, O., S. Jusoh et N. Norwawi. 2012, «Handling ambiguity problems of natural language interface for question answering. International Journal of Computer Science Issues (IJCSI), vol. 9(3), IJCSI Publications, p. 245–265.

Ali, I. (2012). Application of a mining algorithm to finding frequent patterns in a text corpus : A case study of the Arabic. , International Journal of Software Engineering and Its Applications 6(3), 127–134.

Al-Yahya, M., & Al-Khalifa, H. (2010). An ontological model for representing semantic lexicons: An application on time nouns in the Holy Quran. The Arabian Journal for Science and Engineering.

Amit Singhal. (2012). Introducing the Knowledge Graph: things, not strings. Retrieved 20 August 2013 from http://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

Amsler, R. A. (1984). Lexical Knowledge Bases. Proceedings of COLING-84. Stanford University, Stanford, CA, July, pp. 458- 459.

Angrosh, M. A., Cranefield, S., & Stanger, N. (2011). Contextual Information Retrieval in Research Articles : Semantic Publishing Tools for the Research Community. Semantic Web – Interoperability, Usability, Applicability.

Balakrishna, M., & Srikanth, M. (2008). Automatic Ontology Creation from Text for National Intelligence Priorities Framework (NIPF).In Proceeding of 3rd International conference for Ontology for the Intelligent Community (OIC), Fairfax, VA, USA, Dec 3-4.

Balakrishna, M., Moldovan, D., Tatu, M., & Olteanu, M. (2013). Semi-Automatic Domain Ontology Creation from Text Resources. Proceedings of the Twenty-Sixth International Florida Artificial Intelligence Research Society Conference, 3187–3194.

Baqai, S., Basharat, A., Khalid, H., Hassan, A., & Zafar, S. (2009). Leveraging Semantic Web Technologies for Standardized Knowledge Modelling and Retrieval from the Holy Quran and Religious Texts. In: ACM Special Interest Group on Artificial Intelligence, Proceedings of the 7th International Conference on Frontiers of Information Technology. ACM, New York.

Barzdins, G., Liepins, E., Veilande, M., & Zviedris, M. (2009). Ontology Enabled Graphical Database Query Tool for End-Users. In Eighth International Baltic Conference on Databases and Information Systems (DB&IS 2008), 105–116.

Belkin, N.J. (1996). Intelligent information retrieval: Whose intelligence Proceedings des 5. Internationalen Sypmosiums für Information swissenschaft (ISI `96)., Konstanz: Universitätsverlag Konstanz, 25–31.

Bellavia, G., Maio, D., & Rizzi, S. (1992). Resolving the query inference problem by optimizing query formulation cost, 1–30. Technical Report CIOC-C.N.R. n. 85.

Bettina Fazzinga and Thomas Lukasiewicz. (2010). Semantic search on the Web. Semantic Web — Interoperability, Usability, Applicability. 89–96.

Bizer, C., Kobilarov, G., Lehmann, J., & Ives, Z. (2007). DBpedia : A Nucleus for a Web of Open Data. In Proceedings of the 6th International Semantic Web Conference. ISWC 2007 + ASWC 2007, Busan, Korea.

Blanco, RoiHalpin, HarryHerzig, Daniel M.Mika, PeterPound, JeffreyThompson, Henry S.Tran, Thanh. (2013). Web Semantics: Science, Services and Agents on the World Wide Web. Journal of Web Semantics.

Bo, K. (2002). Visual Interfaces for Semantic Information Retrieval and Browsing. In V.

Boldi, P., Bonchi, F., Castillo, C., & Vigna, S. (2009). From "Dango" to "Japanese Cakes": Query Reformulation Models and Patterns. 2009 IEEE/WIC/ACM

International Joint Conference on Web Intelligence and Intelligent Agent Technology, 183–190.

Bontas, E. P., Mochol, M., & Tolksdorf, R. (2005.). Case Studies on Ontology Reuse. In Proceeding of the 5th International Conference on Knowledge Management IKNOW05.

Bratsas, C., Quaresma, P., Pangalos, G., & Maglaveras, N. (2004). Using ontologies to build a knowledge base of cardiology problems and algorithms. Computers in Cardiology. 609–612.

Brill. Eric. (1995).Penn Treebank Tagger, Copyright by M.I.T and University of Pennsylvania.

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. ACM Computing Surveys, 44(1), 1–50.

Castells, P., Fernández, M., & Vallet, D. (2007). An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval IEEE Trans.Knowl. Data Eng., vol. 19, no. 2, pp. 261–272, Feb. 2007.

Chen, Z., & Zhu, Q. (1998). Query construction for user-guided knowledge discovery in databases. Information Sciences, 109(1-4), 49–64.

Ciccarese, P., Ocana, M., Garcia Castro, L. J., Das, S., & Clark, T. (2011). An open annotation ontology for science on web 3.0. Journal of Biomedical Semantics, 2041-1480-2-S2-S4.

Cimiano, P., Haase, P., Heizmann, J., Mantel, M., & Studer, R. (2008). Towards portable natural language interfaces to knowledge bases – The case of the ORAKEL system. Data & Knowledge Engineering, 65(2), 325–354.

Concept, (August), 333-337. Salton, G. and Buckley, C. Improving Retrieval Performance by Relevance Feedback. Journal of theAmerican Society for Information Science, 41,4 (1990),pp. 288-297.

Croft W. Bruce. (1986) .User-Specified Domain Knowledge for Document Retrieval. In Proceedings of SIGIR, pages 201.

Croft, W.B., Turtle, H., and Lewis, D. (1991). The use of phrases and structured queries in information retrieval. SIGIR-91: Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 32-45, New York: ACM.

Damljanovi, D. D. (2011). Natural Language Interfaces to Conceptual Models, PhD thesis, University of Sheffield, UK.

Damljanovic, D., Agatonovic, M., & Cunningham, H. (2010). Natural Language Interfaces to Ontologies : Combining Syntactic Analysis and Ontology-Based Lookup through the User Interaction. In Proceedings of the 7th Extended

Semantic Web Conference (ESWC 2010), Heraklion, Greece, May 31-June 3, 2010. Springer.

Damljanovic, D.; Agatonovic, M.; and Cunningham, H. 2012. FREyA: an Interactive Way of Querying Linked Data using Natural Language. The Semantic Web.125–138. Springer

Damova M., D. Dannelles, R. Enache, M. Mateva, and A. Ranta. Natural language interaction with semantic web knowledge bases and lod. In Towards the Multilingual Semantic Web. Springer, 2013.

Decker, S., Sintek, M., Billig, A., Henze, N., Harth, A., Leicher, A.Neumann, G. (2006). TRIPLE - an RDF Rule Language with Context and Use Cases. Proceedings of the W3C Workshop on Rule Languages for Interoperability. Washington DC. USA.

Deines, I., & Krechel, D. (2013). A German Natural Language Interface for Semantic Search, Semantic Technology. In proceeding of Second Joint International Conference, JIST 2012Nara, Japan, December 2012.

Donderler M, E. S¸aykol, O¨. Ulusoy, U. Gu¨du¨kbay, BilVideo: a video database management system. IEEE Multimedia, 10 (1) (2003) 66–70

Dongilli, P., Franconi, E. (2006). An Intelligent Query Interface with Natural Language Support. In: Proceedings of the 19th International FLAIRS Conference FLAIRS-2006, Melbourne Beach, Florida, May 2006.

Dongilli, P., Franconi, E. And Tessaris, S. (2004). Semantics driven support for query formulation. In International Workshop on Description Logics, Whistler, BC, Canada.

Dongyi Guan. 2013. Structured Query Formulation and Result organization For Session Search, MSc Thesis, Georgetown University.

Dredze, M., McNamee, P., Rao, D., Gerber, A. & Finin, T. (2010). Entity Disambiguation for Knowledge Base Population. In: Proceedings of COLING.

Dror Judith, Dudu Shaharabani, Rafi Talmon and Shuly Wintner. (2004). Morphological Analysis of the Qur'an. Literary and Linguistic Computing, 19(4):431-452, 2004 Dukes,

Dukes Kais and N. Habash. Morphological. (2010). Annotation of Quranic Arabic. Language Resources and Evaluation Conference (LREC).Valletta, Malta, 2010.

Dukes Kais, Atwell Eric and A. M. Sharaf. (2010). Syntactic Annotation Guidelines for the Quranic Arabic Treebank. Language Resources and Evaluation Conference (LREC), Valletta, Malta, 2010.

Dukes Kais, Eric Atwell and Nizar Habash. (2013). Supervised Collaboration for Syntactic Annotation of Quranic Arabic. Language Resources and Evaluation

Journal (LREJ): Special Issue on Collaboratively Constructed Language Resources, 47:1 (33-62).

Dukes, K., Atwell. University of Leeds Quran ontology. http://corpus.quran.com/ontology.jsp. Retrieved 23 September 2010.

Enikuomehin A.O., Okwefulueze D.O. (20120. An Algorithm for Solving Natural Language Query Execution Problems on Relational Databases. International Journal of Advanced Computer Science and Applications 169-175

Erdmann M., A. Maedche, H. Schnurr, and S. Staab. (2000). From manual to semi-automatic semantic annotation: About ontology-based text annotation tools. In, Proceedings of the COLING 2000 Workshop on Semantic Annotation and Intelligent Content, August 2000.

Ferrández, Ó., Izquierdo, R., Ferrández, S., & Vicedo, J. L. (2009). Addressing ontology-based question answering with collections of user queries. Information Processing & Management 45(2), 175–188.

Finin Tim, James Mayfield, Clay Fink, Anupam Joshi, and R. Scott Cost. (2005). Information Retrieval and the Semantic Web. In Proceedings of the 38thInternational Conference on System Sciences.

Finin, T., Mayfield, J., Joshi, a., Cost, R. S., & Fink, C. (1004). Information Retrieval and the Semantic Web. Proceedings of the 38th Annual Hawaii International Conference on System Sciences 113a–113a.

Fuhr, N. (2003). Information Retrieval Methods for Literary Texts.

G Kück, (2004), "Tim Berners-Lee's Semantic Web". South African Journal of Information Management.

G. Leech, R. Garside, E. Atwell. (1983). The Automatic Grammatical Tagging Of The Lob Corpus. Icame Journal Of The International Computer Archive Of Modern English Vol.7.

G. Marcos, A. Illarramendi, I.G. Olaizola, and J. Flórez. (2011). A middleware to enhance current multimedia retrieval systems with content-based functionalities. In: Multimedia systems 17.2 (2011), 149–164.

Gauch Groppe, J., Groppe, S., Ebers, S., & Linnemann, V. (2009). Efficient processing of SPARQL joins in memory by dynamically restricting triple patterns. Proceedings of the 2009 ACM Symposium on Applied Computing - SAC '09, 1231.

GeneInfoViz. Gene Ontology. University Montpellier. Retrieved 6 October 2011 from http://irb.chu-montpellier.fr/fr/PDF/Bioinfo2007.

Geroimenko & C. Chen (Eds.), Visualizing the Semantic Web (1st ed., pp. 99-115): Springer.

Fausto Giunchiglia, Biswanath Dutta and Vincenzo Maltese (2009). Faceted Lightweight Ontologies. Conceptual Modelling: Foundations and applications. LNCS, 36-51. Springer, Heidelberg.

Guha, R., McCool, R., and Miller, E. (2003). Semantic Search. Proceedings of the WWW2003, Budapest, 2003.

Gwizdka, J., & Chignell, M. (1999). Towards information retrieval measures for evaluation of Web search engines. Retrieved 21 December 2012 from http://www.imedia.mie.utoronto.ca/~jacekg/pubs/webIR_eval1_99.pdf.

Habernal, I., & Konopík, M. (2013). SWSNL : Semantic Web Search Using Natural Language. Expert Systems with Applications 40, 3649–3664.

Heflin, J., Hendler, J., & Luke, S. (2001).SHOE: A prototype language for the semantic web. Link"oping Electronic Articles in Computer and Information.

Hersh WR, Haynes RB. (1991). Evaluation of SAPHIRE: an automated approach to indexing and retrieving medical literature. Proceedings 15th Ann Symp Comp Applzc Med Care 1991; 15:808–812

Hitzler P., M. Krotzsch, and S. Rudolph. (2009). Foundations of Semantic Web Technologies. Chapman &Hall/CRC, August 2009.

Hliaoutakis, A., Varelas, G., Voutsakis, E., Petrakis, E.G.M., & Milios, E. (2006). Information retrieval by semantic similarity. International Journal on Semantic Web and Information Systems, 2(3), 55–73.

Hu, Y. (2004). The Semantic Web : Current Status and Future Directions. [Power Point slides]. Retrieved from http://www.cs.nccu.edu.tw/~jong/pub/mis0601talk.pdf.

Huang, J., & Efthimiadis, E. N. (2009). Analysing and evaluating query reformulation strategies in web search logs. Proceeding of the 18th ACM Conference on Information and Knowledge Management - CIKM '09, 77.

Ingason, A. K., Helgadóttir, S., & Loftsson, H. (2008). A Mixed Method Lemmatization Algorithm Using a Hierarchy of Linguistic Identities (HOLI). Advances in Natural Language Processing 205–216, 2008.

Jansen, B.J., Booth, D.L., & Spink, A. (2009). Patterns of query modification during Web searching. Journal of the American Society for Information Science and Technology 60(7), 1358–1371.

Jarar Mustafa., Khan, S., & Latif, K. (2008). Ontology based semantic information retrieval. 2008 4th International IEEE Conference Intelligent Systems, 22–14–22–19. doi:10.1109/IS.2008.4670473

Jarrar Mustafa, Marios D. Dikaiakos. (2012). A Query Formulation Language for the Data Web," IEEE Transactions on Knowledge and Data Engineering 783-798.

Jin, H., Ning, X., Chen, H., & Yin, Z. (2006). Efficient query routing for information retrieval in semantic overlays. Proceedings of the 2006 ACM Symposium on Applied Computing - SAC '06, 1669.

Kassim, J.M., Rahmany, M. (2009). Introduction to semantic search engine. In: International Conference on Electrical Engineering and Informatics, 380–386. Selangor, Malaysia (2009)

Kaufmann, E., Bernstein, A., Zumstein, R. (2006). Querix: A natural language interface to query ontologies based on clarification dialogs. In: ISWC 2006.LNCS, 980–981. Springer, Heidelberg.

Keller, F. (2000). Connectionist and Statistical Language Processing Clustering vs. Classification Supervised vs. Unsupervised Learning.[Power Point slides]. Retrieved from http://www.coli.uni-saarland.de/~crocker/Teaching/Connectionist/lecture13_4up.pdf

Khan, H., Saqlain, S., Shoaib, M., & Sher, M. (2013). Ontology based semantic search in Holy Quran. International Journal of Future Computer and Communication, V2.229.

Kharbat, F., & El-ghalayini, H. (2008). Building Ontology from Knowledge Base Systems. Data Mining in Medical and Biological Research, Book edited by: Eugenia G. Giannopoulos, ISBN 978-953-7619-30-5, pp. 320, December 2008, I-Tech, Vienna, Austria.

Kharkevich Uladzimir, (2010).Concept Search: Semantics Enabled Information Retrieval PhD Thesis, Universit`a degli Studi di Trento.

Kharlamov, E., Giese, M., Soylu, A., Zheleznyakov, D., Bagosi, T., Console, M.,Waaler, A. (2013). Optique 1. 0 : Semantic Access to Big Data the Case of Norwegian Petroleum Directorate's Fact Pages. The 12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia (1), 1–4.

Kumar, C. A., Gupta, A., Batool, M., & Trehan, S. (2006). Latent Semantic Indexing-Based Intelligent Information Retrievals. Journal of Computing and Information Technology - CIT 14 P, 191–196.

L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. C. Doshi,and J. Sachs. (2004. Swoogle: A search and metadata engine for the semantic web. In Proceedings of the 13th ACM Conf. on Information and Knowledge Management, 2004

Lawson, T. (2004) A Conception of Ontology, University of Cambridge, Faculty of Economics, Cambridge.

Lehmann, J., & Lorenz, B. (2011). AutoSPARQL : Let Users Query Your Knowledge Base. In Proceedings of ESWC 1–15.

Lei, Y., Uren, V., Motta, E. (2006).Semsearch: A search engine for the semantic web. In: Staab, S.,Svátek, V. (eds.) EKAW 2006. LNCS (LNAI), vol. 4248, pp. 238–245. Springer, Heidelberg.

Li, Y., Yang, H., Jagadish, H. (2005).NaLIX: An interactive natural language interface for querying xml. In: Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, pp. 900–902. ACM Press, New York

Liu S., F. Liu, C. Yu, and W. Meng. (2004). An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In SIGIR, 2004.

Lopez, V., Fernández, M., Stieler, N., Motta, E., Hall, W., Mkaa, M. K., & Kingdom, U. (2011). PowerAqua : supporting users in querying and exploring the Semantic Web content. Semantic Web Journal.

Lopez, V., Motta, E., Uren, V. and Pasin, M. (2007). AquaLog: An ontology-driven Question Answering System for Semantic intranets. Journal of Web Semantics 5 (2).

Madhu, G., Govardhan, a, & Rajinikanth, T. K. V. (2011). Intelligent semantic web search engines: A brief survey. International Journal of Web & Semantic Technology, 2(1), 34–42. doi:10.5121/ijwest.2011.2103

Maron, M. E. And Kuhns, J. L. 1960. On relevance, probabilistic indexing and information retrieval. J. ACM7, 3, 216–244.

Mitra M., A. Singhal, and C. Buckley. (1998). Improving automatic query expansion. In Proceedings of the 21stAnnual International ACM-SIGIR Conference on Research and Development in Information Retrieval 214, 1998.

Mitra, M., Singhal, A., And Buckley, C. 1998. Improving automatic query expansion. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '98, Melbourne, Australia, Aug. 24–28)

Moldovan. D and Mihalcea. R. (2000). Using Word- Net and lexical operators to improve Internet searches. IEEE Internet Computing, 4(1):34-- 43.

Moore, M., & Eb, S. E. W. (2011). The Semantic Web : An introduction for information professionals. Retrieved from http://innotecture.files.wordpress.com/2011/04/olc-april-2011_moore_semantic-web.pdf

Mourtaga, E., Sharieh, A., & Abdallah, M. (2007). Speaker-Independent Qur'an ic Recognizer Based on Maximum Likelihood Linear Regression. PWASET, 2007.

Muchemi, L. (2007). Swahili Natural Language Statements for Database Querying, 50–58.

Nassourou, M. (2011). Assisting Analysis and Understanding of Quran Search Results with Interactive Scatter Plots and Tables. Retrieved from http://opus.bibliothek.uni-wuerzburg.de/volltexte/2011/5584/

Nassourou, M. (2012). Using Machine Learning Algorithms for Categorizing Quranic Chapters by Major Phases of Prophet Mohammad's Messenger ship, 2(11), 863–871.

Navigli, R. (2009). Word sense disambiguation. ACM Computing Surveys, 41(2), 1–69.

Nilsson, N. J. (1998). Introduction .to Machine Learning. Robotics Laboratory Department of Computer Science, Stanford University.

Noordin, M.F., & Othman, R. (2006). An information retrieval system for Qur'anic texts: A proposed system design. IEEE International Conference on Information and Communication Technologies, ICTTA '06 1704 - 1709.

Nurfadhlina Mohd Sharef, And Shahrul Azman Noah. (2013). Natural Language Query Translation For Semantic Search. International Journal Of Digital Content Technology And Its Application.

OWL Web Ontology Language. (2004) Retrieved 20 august 2012 from http://www.w3.org/TR/owl-ref/.

Palmer. D, Tokenisation and sentence segmentation (2000), Handbook of Natural Language Processing, Marcel Dekker, New York, 2000, pp. 11–36.

Patel-Schneider, P. F. (2005): A Revised Architecture for the Semantic Web Reasoning. In: Proceedings of PPSWR'05, Dagstuhl, Germany.

Ponzetto S.P, R. Navigli, Knowledge-rich word sense disambiguation rivaling supervised system, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL 2010, Uppsala, Sweden, pp. 1522–1531

Popov, B., Kiryakov, A., Kirilov, A., Manov, D., Ognyanoff, D. and Goranov, M. (2003). KIM - Semantic Annotation Platform. In 2nd International Semantic Web Conference, Florida, USA, 2003 834-849.

Qurat and Amma. (2011). Ontology driven Information Extraction from the Holy Qur'an related Documents. 26th IEEEP Students' Seminar 2011 Pakistan Navy Engineering College National University of Sciences &Technology, Pakistan.

Ramanathan Guha, Rob McCool, and Eric Miller. Semantic Search. In WWW '03: Proceedings of the 12th international conference on World Wide Web, New York, NY, USA, 2003. ACM. ISBN1-58113-680-3.

Richardson, M., & Domingos, P. (2003). Building large knowledge bases by mass collaboration. Proceedings of the International Conference on Knowledge Capture - K-CAP '03, 129. doi:10.1145/945649.945665

Rindflesch, T. C., & Aronson, a R. (1993). Semantic processing in information retrieval. Proceedings of the 17th Annual SCAMC, 611-615.

Saad, S., & Salim, N. (2008). Methodology of Ontology Extraction for Islamic Knowledge Text. Postgraduate Annual Research Seminar. Retrieved from http://comp.utm.my/pars/files/2013/04/Methodology-of-Ontology-Extraction-for-Islamic-Knowledge-Text.pdf

Saad, S., Salim, N., & Zainal, H. (2009). Pattern Extraction For Islamic. International Conference on Electrical Engineering and Informatics 5-7 August 2009, Selangor, Malaysia

Sanderson, M. (1994). Word sense disambiguation and information retrieval. In Proceedings of the Special Interest Group on Information Retrieval. SIGIR, Dublin, Ireland. 142–151.

Sanderson, M., Retrieving with good sense. (2000). Information Ret., 2(1): pp. 49-69,

Santos, R. L. T., Macdonald, C., & Ounis, I. (2010). Exploiting query reformulations for web search result diversification. Proceedings of the 19th International Conference on World Wide Web - WWW '10, 881.

Schiff, J. (2011). Semantic Web Technologies Effects on Information Retrieval in Digital Libraries : An Annotated Bibliography.

Seddah, D. (2010). Lemmatization and lexicalized statistical parsing of morphologically rich languages : The case of French. In: SPMRL 2010 - 1st Workshop on Statistical Parsing of Morphologically-Rich Languages at NAACL HLT 2010, 5 June 2010, Los Angeles, CA, USA.

Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The semantic web revisited. IEEE Intelligent Systems, 21(3), 96–101. doi:10.1109/MIS.2006.62

Shahzadi, N., & Shaheen, A. (2011). Semantic network based semantic search of religious repository. International Journal of Computer Applications 0975 – 8887.

Sharaf, A. M., & Atwell, E. S. (2011). QurAna : Corpus of the Quran annotated with Pronominal Anaphora. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA). 2012.

Sharaf, A. M., & Atwell, E. S. (2012). QurSim : A corpus for evaluation of relatedness in short texts. Proceedings of the Eight International Conference on

Language Resources and Evaluation (LREC'12). European Language Resources Association (ELRA). 2012.

Sharaf, A., & Atwell, E. (2009). Knowledge representation of the Quran through frame semantics A corpus-based approach. Proceeding Of The Fifth Corpus Linguistic Conference, Liverpool, UK.

Sherif, M. A., & Ngomo, A. N. (2013). Semantic Quran A Multilingual Resource For Natural-Language Processing. Semantic Web Journal.

Shinsuke Mori, Tetsuro Sasada, Yoko Yamakata, and Koichiro Yoshino. (2012). A machine learning approach to recipe text processing. In Proceedings of Cooking with Computer workshop.

Shobana R., D.Venkatesan (2012). FFQI-Fast Formulation Query Interface For. Journal of Theoretical and Applied Information Technology 37(1), 125–131.

Shobana R., D.Venkatesan. (2012). Ffqi-Fast Formulation Query Interface for Databases. Journal of Theoretical and Applied Information Technology.

Singh, Y., Bhatia, K. & Sangwan, O. (2007). A Review of Studies in Machine Learning Techniques. International Journal of Computer Science and Security, 1(1), 70-84.

Sivic, J., & Zisserman, A. (2003). Video Google: a text retrieval approach to object matching in videos. Proceedings Ninth IEEE International Conference on Computer Vision, (Iccv), 1470–1477 vol.2. doi:10.1109/ICCV.2003.1238663

Smith, C., & June, S. (2005). Latent Semantic Indexing : Advancing the model A Review of the Research. Info 622 - Content Representation, 2005.

Solskinnsbakk, G. (2012). Contextual Semantic Search Navigation, PhD Thesis, Norwegian University of Science and Technology.

Stratica, N., Kosseim, L., & Desai, B. C. (2005). Using semantic templates for a natural language interface to the CINDI virtual library. Data & Knowledge Engineering, 55(1), 4–19. doi:10.1016/j.datak.2004.12.002

Stubinz, J., & Whighli, S. (n.d.). Information Retrieval System Design for Very High Effectiveness. Retrieved from http://goanna.cs.rmit.edu.au/~jz/sci/p3.pdf.

Tablan, V., Damljanovic, D., & Bontcheva, K. (2008). A Natural Language Query Interface to Structured Information, In ESWC 2010, volume 6088 of LNCS, pages 106{120. Springer, 2010 .

Tannier, X., Girardot, J., Mathieu, M., & Saint-étienne, F.-. (n.d.). Natural Language Queries for Information Retrieval in Structured Documents. Retrieved from http://perso.limsi.fr/xtannier/Publications/files/Tannier_AISTA04.pdf.

Tao, C., Embley, D. W., & Liddle, S. W. (2009). FOCIH : Form-based Ontology Creation and Information Harvesting. In Proceedings of the 28th International Conference on Conceptual Modeling , Gramado, Brazil.

Tauberer, J. (2008). Why we need a new standard for the Semantic Web, (October 2005). Retrieved from http://www.rdfabout.com/intro/?section=1

Thabet, N. (2004). Stemming the Qur'an. Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages - Semitic '04, 85.

Unger, C., Bühmann, L., Lehmann, J., Ngonga Ngomo, A.-C., Gerber, D., & Cimiano, P. (2012). Template-based question answering over RDF data. Proceedings of the 21st International Conference on World Wide Web - WWW '12, 639.

Voorhees, M., & Nj, R. (n.d.). Expansion using Lexical-semantic. Proceeding of ACM SIGIR International Conference on Research on Research and Development in Information retrieval 61-69.

Wang, C., Xiong, M., Zhou, Q., & Yu, Y. (2007). PANTO : A Portable Natural Language Interface to Ontologies. Proceedings of the Fourth European Semantic Web Conference 473-487.

Woods W.A. (1997): Conceptual Indexing: A Better Way to Organize Knowledge. A Sun Labs Technical Report: TR-97-61.Editor, Technical Reports, 901 San Antonio Road, Palo Alto, California 94303, USA

Woods William. (1997). Conceptual indexing: A better way to organize knowledge. Technical Report TR-97-61, Sun Microsystems Laboratories, USA, 1997.

Xu J., Croft W.B (1996): Query Expansion Using Local and Global Document Analysis. In the Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR 96), Zurich, Switzerland.

Yahya, M., Berberich, K., & Elbassuoni, S. (2011). Natural Language Questions for the Web of Data. In Proceedings of the 2012 joint conference for Empirical methods of Natural Language Processing and Computational Natural Language Learning.

Yonggang Qiu, H.P. Frei, (1993). Concept Based Query Expansion.SIGIR'93, 160-169.

Zenz, G., Zhou, X., Minack, E., Siberski, W., & Nejdl, W. (2009). From keywords to semantic queries—Incremental query construction on the semantic web. Web Semantics: Science, Services and Agents on the World Wide Web, 166–176.

Zhai, C., (1997). Fast statistical parsing of noun phrases for document indexing. In 5th conference on applied natural language processing (ANLP-97) (pp. 312–319).

Zou, Y., Finin, T., & Chen H. (2004). F-OWL: An inference engine for the semantic web, Proceedings of the Third International Workshop (FAABS), April 16–18, 2004 , USA.