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