UNIVERSITI PUTRA MALAYSIA

A NEW APPROACH FOR INSTANCE-BASED SCHEMA MATCHING

OSAMAH ABDUL SATTAR MAHDI

FSKTM 2014 5

A NEW APPROACH FOR INSTANCE-BASED SCHEMA MATCHING

By

OSAMAH ABDUL SATTAR MAHDI

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Master of Science

May 2014

بســـم الله الرحمن الرحيم

﴿ كَمَا أَرْسَلْنَا فِيكُمْ رَسُولاً مِّنكُمْ يَتْلُو عَلَيْكُمْ آيَاتِنَا وَيُزَكِّيكُمْ وَيُعَلِّمُكُمُ الْكِتَابَ وَالْحِكْمَةَ وَيُعَلِّمُكُم مَّا لَمْ تَكُونُواْ تَعْلَمُونَ ﴾

سورة البقرة - آية 151

DEDICATION

This thesis is dedicated to my Dearest, precious and First Teachers:

My Father and Mother
I will always be grateful for your endless love, unlimited support and deep faith in me

And
My brother and sisters, who are like candles that burn to provide others light,
Thanks to Allah for sending these angels to my world.

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfilment of the requirement for the degree of Master of Science

A NEW APPROACH FOR INSTANCE-BASED SCHEMA MATCHING

By

OSAMAH ABDUL SATTAR MAHDI

May 2014

Chairman: Professor Hamidah Ibrahim, PhD

Faculty : Computer Science and Information Technology

Schema matching is a crucial phase in data integration that aims to find correspondences between schema attributes by utilizing schema information. However, this information is not always available or useful to be used since it could be abbreviation. Consequently, instances could be an alternative choice for schema information. Various instance based schema matching approaches have been proposed to achieve the goal of discovering correspondences between schema attributes, by treating the instances as strings including the numeric instances. This prevents discovering common patterns or performing statistical computation among the numeric instances. As a consequence, this causes unidentified matches especially for attribute with numeric instances which further reduces the quality of match results.

This thesis aims at proposing an efficient approach which is able to identify attribute matches between schemas by fully exploiting the instances. The approach utilizes the concept of pattern recognition to determine attribute matches for numeric and mix instances. This is acquired by automatically creating regular expression based on the instances. While, for alphabetic instances the approach calculates the semantic similarity score by utilizing Google similarity to capture the semantic relationships between instances. The proposed approach consists of five main phases, namely: (i) analysing instances, (ii) classifying schema attributes, (iii) extracting the optimal sample size, (iv) identifying instance similarity, and (v) identifying the match.

Three analyses have been designed and conducted on two different data sets, namely: (i) Restaurant and (ii) Census, with respect to precision (*P*), recall (*R*), and F-measure (*F*). The first analysis aims at identifying the optimal sample size of tuples to be used during the phase of extracting the optimal sample size. The purpose of identifying the optimal sample size is to reduce the number of comparisons between the instances which lead to reduce the processing time of matching operation. This analysis showed that the optimal sample size is 50% from the actual table size of both data sets. The second analysis aims to investigate and to prove that combining both Google similarity and regular expression as in our proposed approach achieve higher accuracy compared to utilizing Google

similarity or regular expression separately. The results showed that our proposed approach achieved precision (*P*), recall (*R*), and F-measure (*F*) in the range of 93% - 99% for both data sets. On the other hand, Google similarity and regular expression which are performed separately achieved precision (*P*), recall (*R*), and F-measure (*F*) in the range of 36% - 74%. While the third analysis intents to compare the performance of our proposed approach to the previous approaches. The results showed that our proposed approach outperformed the previous approaches although only a sample of instances is used instead of considering the whole instances during the process of instance based schema matching as used in the previous works.

Abstrak tesis dipersembahkan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan mendapatkan Ijazah Sarjana Sains

PENDEKATAN BAHARU UNTUK PEMADANAN SKEMA BERASASKAN KETIKAAN

Oleh

OSAMAH ABDUL SATTAR MAHDI

Mei 2014

Pengerusi : Profesor Hamidah Ibrahim, PhD

Fakulti : Sains Komputer dan Teknologi Maklumat

Pemadanan skema adalah fasa penting dalam integrasi data yang bertujuan untuk mencari koresponden antara atribut skema dengan menggunakan maklumat skema. Walau bagaimanapun, maklumat ini tidak selalunya tersedia atau berguna untuk digunakan kerana ia boleh jadi singkatan. Akibatnya, ketikaan boleh jadi satu pilihan alternatif bagi maklumat skema. Pelbagai pendekatan padanan skema berasaskan ketikaan telah dicadangkan untuk mencapai matlamat dalam penemuan koresponden antara atribut skema, dengan menganggap ketikaan sebagai rentetan termasuklah ketikaan numerik. Ini menghalang penemuan pola biasa atau menjalankan pengiraan statistikal di kalangan ketikaan numerik. Sebagai akibat, ini menyebabkan pemadanan yang tidak dikenalpasti terutamanya bagi atribut dengan ketikaan numerik yang seterusnya mengurangkan kualiti hasil pemadanan.

Tesis ini bertujuan mencadangkan satu pendekatan efisien yang dapat mengenal pasti padanan atribut antara skema dengan mengeksploitasi sepenuhnya ketikaan. Pendekatan ini menggunakan konsep pengecaman pola untuk menentukan pemadanan atribut bagi ketikaan numerik dan campuran. Ini diperolehi dengan mencipta ungkapan biasa secara automatik berdasarkan kepada ketikaan. Sementara itu, bagi ketikaan abjad pendekatan ini mengira skor persamaan semantik dengan menggunakan persamaan Google bagi mendapatkan pertalian semantik antara ketikaan. Pendekatan yang dicadangkan mengandungi lima fasa utama, iaitu: (i) menganalisis ketikaan, (ii) mengelaskan atribut skema, (iii) mengekstrak saiz sampel yang optimal, (iv) mengenal pasti persamaan ketikaan, dan (v) mengenal pasti pemadanan.

Tiga analisis telah direka bentuk dan dijalankan ke atas dua set data yang berbeza, iaitu: (i) Restoran dan (ii) Census dengan merujuk kepada *precision* (*P*), *recall* (*R*), dan *F-measure* (*F*). Analisis pertama bertujuan untuk mengenal pasti saiz sampel tuple yang optimum untuk digunakan semasa fasa mengekstrak saiz sampel yang optimum. Tujuan mengenal pasti saiz sampel yang optimum adalah untuk mengurangkan bilangan perbandingan antara ketikaan yang mana dapat mengurangkan masa pemprosesan

operasi pemadanan. Analisis ini menunjukkan bahawa saiz sampel yang optimum adalah 50% daripada saiz jadual yang sebenar bagi kedua-dua data set. Analisis kedua bertujuan untuk menyiasat dan membuktikan bahawa menggabungkan kedua-dua persamaan Google dan ungkapan biasa sebagaimana dalam pendekatan cadangan kami mencapai ketepatan yang lebih tinggi berbanding menggunakan persamaan Google atau ungkapan biasa secara berasingan. Hasil menunjukkan bahawa pendekatan cadangan kami mencapai *precision* (*P*), *recall* (*R*), and *F-measure* (*F*) dalam julat 93% - 99% untuk kedua-dua set data. Sebaliknya, persamaan Google dan ungkapan biasa yang dilaksanakan secara berasingan mencapai *precision* (*P*), *recall* (*R*), dan *F-measure* (*F*) dalam julat 36% - 74%. Manakala analisis ketiga bertujuan untuk membandingkan prestasi pendekatan cadangan kami dengan pendekatan yang sebelum ini. Hasil menunjukkan bahawa pendekatan cadangan kami mengatasi pendekatan sebelumnya walaupun suatu sampel ketikaan digunakan daripada mempertimbangakan keseluruhan ketikaan semasa proses padanan skema berasaskan ketikaan sebagaimana digunakan dalam kajian sebelum ini.

ACKNOWLEDGEMENTS

In the name of *ALLAH*, the most merciful and most compassionate. Praise to *ALLAH* S. W. T. who granted me strength, courage, patience and inspirations to complete this work.

In the first place I would like to record my deepest gratitude to my supervisor Professor Dr. Hamidah Ibrahim, for her continuous and caring help, advice, and encouragement during my Ms.c. studies at the Universiti Putra Malaysia. I thank her for being very patient with my questions and my progress, for the countless lessons on writing and presenting technical materials, on doing research in general, and also for the English lessons, even though she is not a language teacher. I gratefully acknowledge Associate Professor Dr. Lilly Suriani Affendey for her advices, supervision, and crucial contribution, which made the backbone of this research and so to this thesis.

Where would I be without my family? My parents deserve special mention for their inseparable support and prayers. My Father, Professor Dr. Abddul Sattar, in the first place is the person who put the fundament my learning character, showing me the joy of intellectual pursuit ever since I was child. My Mother is the one who sincerely raised me with her caring and gently love. Aseel, Nawfal, and Zena thanks for being supportive and caring siblings.

Osamah Abdul Sattar Mahdi

2014

This thesis was submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfillment of the requirement for the degree of Master of Science. The members of the Supervisory Committee were as follows:

Hamidah Ibrahim, PhD
Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Lilly Suriani Affendey, 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, illustrations and citations have been duly referenced;
- this thesis has not been submitted previously or concurrently for any other degree at any other institutions;
- intellectual property from the thesis and copyright of thesis are fully-owned by Universiti Putra Malaysia, as according to the Universiti Putra Malaysia (Research) Rules 2012;
- written permission must be obtained from supervisor and the office of Deputy Vice-Chancellor (Research and Innovation) before thesis is published (in the form of written, printed or in electronic form) including books, journals, modules, proceedings, 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 plagiarisim detection software.

Signature: _____ Date: _____

Name and Matric No.:_____

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 Universiti Putra Malaysia (Graduate Studies) Rules 2003 (Revision 2012-2013) are adhered to.

Signature: _____	Signature: _____
Name of Chairperson of Supervisory Committee: _____	Name of Member of Supervisory Committee: _____

TABLE OF CONTENTS