Al-Quds University
Deanship of Graduate Studies
Computer Science Department

Thesis Approval

Using Thesaurus as a Schema Matching Approach at the Element Level

Prepared By: Thabit Sulaiman Odeh Sabbah
Registration No: 20510158

Supervisor: Dr. Rashid Jayousi

Master thesis submitted and accepted. Date: 15 / 6 / 2009
The names and signatures of the examining committee members are as follows:

1- Head of Committee: Dr. Rashid Jayousi Signature: R. Jayousi
2- Internal Examiner: Dr. Labib Arafeh Signature: ……………
3- External Examiner: Dr. Nizar Awartani Signature: ……………

Jerusalem – Palestine

1430 / 2009

Abstract

For more than two decades up to now, schema matching is a basic problem in many application domains such as data integration, Electronic business, data warehousing, and semantic query processing. Many individual and hybrid approaches were proposed to solve this problem semi-automatically either at element or schema levels.

In this study we propose a new approach to solve this problem at the element level using thesaurus. Thesaurus which is a well-known tool from the domain of Information Retrieval (IR) is used here to determine the matches between schemas elements through analyzing their textual descriptions.

Our matching process passes through three main steps: 1) data extraction, 2) data analysis and element matching, 3) result viewing. A full application was implemented, and many experiments were carried out to test our methodology. Initially, raw data were collected from the Internet for three different knowledge domains ;(Agriculture, Computer Science, and Education), and then it was processed to become suitable to our experiments. In our experiments we tried to find the equivalent courses from two different lists of courses for each knowledge domain depending on the analysis of course description. The similarity matrix between elements was built and the maximum values were considered as matches. The completeness, effectiveness, and accuracy measures were used to evaluate our results, two experiments of three show positive results with average Precision of 24% and 32% in average for Recall.

الملخص

تعتبر مشكلة مطابقة الهيكليات من المشاكل التي تفتقر إلى حل محوسب فعّال ومتكامل منذ أكثر من عقدين من الزمن، ولا يزال مجال البحث فيها متاحاً. وتنبع أهمية هذه المشكلة باعتبارها الخطوة الأولى والرئيسة في مختلف المجالات التي تبنى بشكل أساسي على عملية تكامل البيانات، كالتجارة الإلكترونية، وبناء الاستعلامات المنطقية وغيرها، ومما يزيد من أهميتها أيضاً التوسع المطرد في المجالين التجاري والعلمي المتزامن مع تطور العلوم التطبيقية والانتشار الواسع للشبكة العنكبوتية وخدماتها، والتي وفرت البيئة المناسبة لعملية تكامل البيانات من مصادر مختلفة إضافة إلى فرض هذا الواقع في أحيان متعددة. وقد طرح العديد من الاقتراحات والآليات لمعالجة هذه المشكلة سواء على مستوى الهيكليات بشكل كامل أو على مستوى عناصرها كل على حده.

في هذه الدراسة اقترحنا نهجاً جديداً لحل هذه المشكلة على مستوى عناصر الهيكليات، يقوم هذا النهج على مبدأ استخدام المُكَنِّز* (Thesaurus) في عملية تحليل النصوص التي تحتوي وصفاً لعناصر الهيكليات، لتحديد العناصر المتشابهة ما بين الهيكليات.

تتم عملية المطابقة المقترحة من خلال ثلاث مراحل: 1) استخراج البيانات، 2) تحليل البيانات ومطابقة العناصر، 3) استعراض النتائج. ولفحص هذه المقترح، فقد تم بناء نظام محوسب متكامل، وتم إجراء العديد من التجارب.

تم جمع بيانات أولية من شبكة الانترنت في ثلاثة مجالات معرفية مختلفة: مجال الزراعة، مجال التعليم، ومجال علم الحاسب، ومن ثم تم معالجتها لتصبح ملائمة لإجراء التجارب عليها. في تجاربنا حاولنا مطابقة المساقات الدراسية بين مجموعتين مختلفتين في نفس المجال المعرفي من خلال تحليل وصف هذه المقررات باستخدام المُكَنِّز، ولتحقيق ذلك فقد تم بناء مصفوفة التشابه بين عناصر المجموعتين، وتم اعتماد أقصى القيم في مصفوفة التشابه لاستخلاص العناصر المتشابهة. وقد تم تقييم نتائج التجارب باستخدام المقاييس المعتمدة في هذا المجال، وكانت النتائج إيجابية في تجربتين من أصل ثلاثة تجارب تم إجراؤها، حيث كان متوسط مستوى الدقة 24% في حين كان متوسط الاسترداد (استدعاء المعلومات) 32%.

[*] يُعرف المكنز وظيفياً على أنه أداة تحكم لفظية تستخدم للترجمة من اللغة الطبيعية للوثائق إلى نظام لغوي مقيد (لغة وثائق ولغة معلومات). أما من حيث التكوين أو البناء، فالمكنز عبارة عن مجموعة مصطلحات متعلقة ببعضها بروابط وعلاقات ترادفية وهرمية واتصالية وتفريعية لتغطي مجالا محددا من مجالات المعرفة (الشامي، 2005).

Table of Contents

Chapter One

Introduction

Data exchange and integration among different systems is one of the famous problems in computer science, this is because these systems were almost developed separately for different uses and their data are stored in variant structures (schemas) and formats. The development of technology and knowledge makes it necessary to exchange and integrate data between such systems. Most of solutions that were introduced to overcome this problem were mainly built on schema mapping. Schema mapping (sometimes it's called "schema matching") is an increasingly important problem itself (Dong Ce, Bailey James, 2006).

Over the past three decades, many approaches were proposed to automate this process with less cost and high accurate and complete outcome. Tools and methods from different domains were used to solve this problem. Thesaurus which is a well-known tool from the domain of Information Retrieval (IR) is proposed here as a new approach to solve the problem of schema matching. The idea of this approach is to discover the matches between schemas through examining the similarity of textual description of schemas' elements using thesaurus. Evaluations of the main arguments (completeness, effectiveness, and accuracy) of the results were measured, and showed good indicators about using thesaurus as schema matcher.

1.1 Motivation

Through our work at The Information and Communication Technology Center (ICTC) at Al-Quds Open University (QOU), the process of exchanging and integrating data between different systems; such as registration, accounting, stock, library, e-learning system… etc., is one of the issues that consume great attention and efforts. Exchanging and integration data not only an issue between internal systems, but also considered as a challenge when it required with external data sources such as Universities, Collages, Ministries, or other related crops. Some of these data integration processes are carried out partially automatically, but other most impose manual intervene, especially when some systems are not developed by the ICTC and doesn't match the global design and naming profiles used by other systems.

This mater motivated me to address this area of data integration; our sense was that an improvement in the ways of work and the quality of outcome could be achieved, since the terms related to the used systems are nearly limited. Through surveying the related literature, it was recognized that the problem was universal, and still an open research area, our ambitious to solve this problem was increased, the goal becomes to participate in solving this world wide problem.

1.2 Organization

Our methodology of schema matching using thesaurus is presented by details throughout this thesis, it can be considered as theoretical and technical specifications of our methodology and its implementation. An Introduction to the subject and the motivation issues were introduced above, the rest of this thesis is organized as follows:

Chapter two presents a literature review about the problem of schema matching, surveys, and the most common approaches that are proposed to solve this problem in time line. Chapter three briefly introduces the thesaurus which is the tool from IR domain that we will use as the core of schema matching process. Our methodology is discussed in details in chapter four, whereas chapter five explains the issues of our implementation. Case studies including both thesauri and schemas data in addition to our experiments and its limitations are expressed in chapter six. In chapter seven we discuss our results and its evaluation. Finally, we conclude our work and explain the future work in chapter eight.

Chapter Eight

Conclusion and Future Work

Many approaches were proposed to solve the problem of schema matching, as indicated in the literature review, we out lined a brief description for some of these approaches in a time line, in addition to introduced surveys. Our study proposes the Thesaurus as a tool to solve this problem. Our methodology was developed depending on the literature of the subject. A full application was developed to test our methodology, and experiments were held on.

The goal was to solve the problem of schema matching by exploiting thesaurus in schema matching through the process of analyzing textual descriptions' of schemas' elements. To achieve this goal, we develop our methodology which consisted of three phases; data extraction, data analysis and elements matching, and the result viewing phase. Through these phases the elements and their descriptions were extracted from the input schemas and filtered against stop words, then thesaurus was applied onto each element's description, a similarity matrix were built and the best matches were chosen and viewed.

To test this methodology we have developed a full application which consisted of two parts; Database and GUI. Our experiments' data were collected from the Internet; three thesauri of different knowledge domains in addition to input schemas of our experiments' were downloaded and justified to become suitable to our experiments. In our experiments we try to find the equivalent courses from two different lists of courses in certain knowledge domain. The course number represents the element name of the schema while course description represents the textual description of the element.

Through our subject literature's study, and while we were designing, developing, implementing, testing our methodology, and evaluating results, all indicators point towards that thesaurus can be used as a tool to solve the problem of schema matching. Our experiments' results were promising; for two of three of our experiments we have a positive result (i.e. we have correct matches), with closed precision and recall ratios. In addition, the following conclusions were arrived:

1. For more than two decades, the problem of schema matching stills an open research topic.
2. The importance of schema matching stems from the increasingly need of it in wide and different areas, for example, and not limited to data warehousing, semantic query processing, data integration, and E business.
3. Many individual and hybrid approaches were developed to solve this problem; Individual approaches that based on single matching criteria can be classified into these categories: Instance vs. schema, Element vs. structure matching, Language vs. constraint, Matching cardinality, and Auxiliary information. Hybrid approaches combine either some of individual approaches into one, or combine the results from many individual approaches into one composite matcher.
4. Many classifications of schema matching approaches were introduced through many surveys. Classifications depend on different levels of opposite factors such as: Instance vs. schema, Element vs. structure matching, Language vs. constraint, Matching cardinality, and others.
5. None of proposed techniques introduce a satisfactory solution of this problem although some of these techniques were better than others.

6. The results of our experiments show that thesaurus which is a well-known tool from the Information Retrieval (IR) domain can be used as a tool to solve the problem of schema matching. The base of this is the process of analyzing texts that describe the schemas and schemas' elements.
7. The speed of searching about terms into thesaurus database is strongly affected by structure design of the database and the used algorithm to search the database.
8. Many factors affect the outcome of schema matching using thesaurus; these factors are related to both thesauri and schemas, such as number of terms in thesauri, and the depth of descriptions of schemas elements. Through our experiments it was noticeable that the average words count in textual descriptions has more effect than the number of terms in thesaurus.
9. More tunings of used thesauri and elements' descriptions are required to achieve better results than we have in our experiments.

8.1 Future Work

This study investigated a new approach to solve the problem of schema matching using thesaurus at the element level. It can be considered as a start point toward a generic schema matching that doesn't depend on single factor to solve the problem; as a result, currently we are developing a generic tool that uses thesaurus as the core of schema matching process in addition to depending on more factors that can be extracted from the schema such as elements' attributes, data types, … etc. The ability of using other schema formats such as text files, ER diagrams, DTDs, and others, as the input of our matcher is one of our considerations within the future development of our methodology. Our matcher can be integrated with other schema matching systems that depend on different criteria.

Our results shows that may be a relation between the number of terms in thesaurus and the average words count in elements' description, the exploration of this relation is one of our future concerns.