جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

École Nationale Supérieure d'Informatique et d'Analyse des Systèmes
Centre d'Études Doctorales en Sciences des Technologies de l'Information et de l'Ingénieur

AVIS DE SOUTENANCE DE THESE DE DOCTORAT

Madame Aola YOUSFI

soutiendra publiquement sa thèse de Doctorat en Informatique

le samedi 04 juin 2022 à 10H00 au Grand amphi à l'ENSIAS

Intitulé de la thèse

hMatcher: A Schema-Driven and Semantic-Aware Holistic Schema Matching Approach

UNIVERSITE MOHAMMED V - RABAT ★ ENSIAS

Devant le Jury composé de :

Président :
Pr. Ali IDRI, PES, ENSIAS, Université Mohammed V de Rabat

Directeur de thèse :
Pr. Ahmed ZELLOU, PES, ENSIAS, Université Mohammed V de Rabat

Co-Directeur de thèse :
Pr. Moulay Hafid EL YAZIDI, PH, FSJES Agdal, Université Mohammed V de Rabat

Rapporteurs :
Pr. El Habib BEN LAHMAR, PES, FSBM, Université Hassan II de Casablanca
Pr. Mohammed BERRADA, PES, ENSAF, Université Sidi Mohamed Ben Abdellah de Fès
Pr. Amal TMIRI, PH, ENSAM, Université Mohammed V de Rabat

Examinateur :
Pr. Hassan SILKAN, PH, Faculté des Sciences, Université Chouaib Doukkali d'El Jadida

hMatcher: A Schema-Driven and Semantic-Aware Holistic Schema Matching Approach

Abstract: There is an abundance of scattered and autonomous data sources, each organizing data in a different manner. The user must then search for the data himself/herself: query each data source separately, which appears to be very tedious, time-consuming and even unfeasible especially if we add to that the fact that there are hundreds, thousands and millions of data sources.

Data integration came to save the day. It suggests to incorporate a uniform software interface between the user and a large number of heterogeneous, autonomous and distributed data sources. This way, the user will send his query only once (instead of querying each source separately) and the data integration system will do the rest. Examples of data integration approaches include but are not limited to common user interface, middleware data integration and application-based integration; choosing one or the other depends on the actual needs. Nonetheless, a data integration system can operate properly only if it solves the problems of interoperability, autonomy, heterogeneity, distribution and scalability. These issues that often pop up when integrating data sources. In order to cope up with these problems, we have to perform schema matching which is very crucial for every data integration system.

Although it is very important, schema matching is very challenging for four main reasons. First, matching schemas requires comparing two schemas at a time, which is not very practical when we wish to match a huge number of schemas. Second, matching two schemas often entails comparing every element in the first schema to every element in the second, which results in a huge search space. Third, hierarchical data structures of different data sources representing the same real-world domain often have different tree representations. They start at a different root element and branch to different leaf elements, which results in a syntactic heterogeneity. Furthermore, node elements' labels do not have a universal naming standard to comply with, which results in a semantic heterogeneity. On top of that, we often do not have precise and complete definitions of schema elements, which makes the schema matching much more complicated.

In this dissertation, we present hMatcher, an unprecedented schema-driven and semantic-aware holistic schema matching approach for hierarchical data structures. hMatcher consists of two key

UNIVERSITÉ MOHAMMED V - RABAT
ENSIAS

modules. First, the pre-processing module aims at clarifying the meanings of schema elements using new enrichment techniques along with various resources, namely an acronyms & abbreviations database and a hierarchical lexical dictionary. Second, the matching module identifies the matches using a holistic matching algorithm and a context-based semantic similarity measure designed to compute the semantic similarity between elements consisting of a single word as well as elements consisting of many words.

Experimental results on six real-world domains show high effectiveness and efficiency of hMatcher. The results also indicate that the semantic similarity measure proposed in this dissertation achieves high accuracy and surpasses the state of the art similarity measures. The findings also show that hMatcher outperforms current matching methods in terms of matching accuracy and search space.

Keywords: Holistic Schema Matching; Schema Elements Frequency; Rare Schema Elements; Matching Accuracy; Search Space; Semantic Similarity; Similarity Measures.

UNIVERSITE MOHAMMED V - RABAT
ENSIAS

Résumé : Hétérogènes, autonomes et distribuées sont les caractéristiques des sources de données; cela dit, leurs schémas de données seront également hétérogènes, autonomes et distribués. Partant de ce principe, il est nécessaire de se demander s'il y a moyen d'accéder à toute sorte de donnée via une interface unique tout en faisant croire à l'utilisateur qu'il y a une seule source de données. Effectivement, ceci est possible grâce à l'intégration de données qui est de plusieurs types : l'intégration par redéveloppement, l'intégration par des middlewares, l'intégration physique, l'intégration virtuelle, l'intégration hybride, l'intégration SOA, etc. Le choix de l'une ou de l'autre approche est fait en fonction des besoins présentés. Cependant, un système d'intégration est considéré performant seulement s'il résout les problèmes d'interopérabilité, d'autonomie, d'hétérogénéité, de répartition et d'évolutivité. Ces problèmes sont souvent rencontrés lors de l'intégration de multiples sources de données. Afin de remédier à ces problèmes, il est très crucial de mettre en place une approche de schema matching.

Il s'avère que la mise en place d'une approche de schema matching est beaucoup plus compliquée que ça en a l'air. En effet, les approches de schema matching rencontrent souvent quatre limitations principales. Premièrement, appliquer une approche de matching entre

différents schémas nécessite une comparaison entre deux schémas à la fois, ce qui n'est pas très pratique lorsque l'on souhaite intégrer un très grand nombre de sources de données. Deuxièmement, effectuer le matching entre deux schémas implique une comparaison entre chaque élément du premier schéma à chaque élément du second, ce qui se traduit par un énorme espace de recherche. Troisièmement, les structures de données hiérarchiques de différentes sources de données représentant le même domaine ont souvent des représentations arborescentes différentes. Elles commencent à un élément racine différent et se ramifient vers différentes feuilles, ce qui entraîne une hétérogénéité syntaxique. De surcroît, les étiquettes des nœuds n'ont pas une règle de nommage universelle à respecter, ce qui entraîne une hétérogénéité sémantique. Quatrièmement, nous manquons souvent de définitions précises et complètes des éléments de schéma, ce qui rend le schema matching beaucoup plus difficile à réaliser.

Dans cette thèse de Doctorat, nous présenterons hMatcher, une nouvelle approche holistique de schema matching conçue particulièrement pour les structures de données hiérarchiques. hMatcher se compose de deux modules clés. Tout d'abord, le module de pré-traitement qui a pour but de clarifier les éléments de schéma en utilisant de nouvelles techniques d'enrichissement ainsi que de diverses ressources, à savoir une base de données d'acronymes & d'abréviations et un dictionnaire lexical. Ensuite, le module de correspondance qui projette d'identifier les correspondances sémantiques à l'aide d'un algorithme de correspondance holistique en plus d'une mesure de similarité sémantique basée principalement sur le contexte et conçue pour calculer la similarité sémantique entre des éléments de schéma constitués d'un seul mot ainsi que des éléments de schéma constitués de plusieurs mots.

Les expériences appliquées à des ensembles de données de six domaines différents montrent que notre approche hMatcher obtient une efficacité et une efficience élevées. Les résultats montrent également que hMatcher surpasse les méthodes de schema matching actuelles en termes de précision et d'espace de recherche. Les résultats indiquent également que la mesure de similarité sémantique proposée dans cette thèse de Doctorat atteint une grande précision et surpasse largement les mesures de similarité actuelles.

Mots clés : Schema matching holistique ; Fréquence des éléments de schéma ; Éléments de schéma rares ; Précision de correspondance sémantique ; Espace de recherche ; Similarité sémantique ; Mesures de similarité.

UNIVERSITE MOHAMMED V RABAT ENSIAS