Andra-Denis Ionescu
47224985
A.D.Ionescu@student.tudelft.nl

An objective evaluation of schema matching techniques
Thesis Proposal

2019-04-07

1 Introduction

Schema matching is an important matter in the database area such as data integration, data warehouse or message mapping. The problem relates to identifying corresponding elements in different schemas, difficult task that used to be performed manually. Nowadays, multiple algorithms and systems have been developed to address this issue, from semantic matching to clustering and machine learning algorithms. In a big setting, where there are hundreds of databases, each with its own schema and terminology, is very challenging to find a matching between any element.

Therefore, it has become an important issue in the database community and it has been actively researched on. Building and maintaining such a system that is spread across multiple databases is a very labour intensive task, that requires human supervision. Most of the times, the relations between elements can not be distinguished easily, mostly because its poor design and documentation. Moreover, with the hardware evolvement and the possibility to store even more data, schema matching importance is increasing tremendously.

The complexity of the matter relies on the possibility to distinguish a match and a non match. The match operation is analysed not only syntactically, but also semantically. Sometimes the corresponding data is analysed in order to disambiguate the elements. Some methodologies are using either the syntax to find similarities, or only the semantics, others have been trying to use unsupervised learning techniques, such as clustering, while others have been focusing on traditional machine learning algorithm and similarities measures.

The existing solutions claim to solve the problem of schema matching, but in an abstract and vague way. They describe in a concise manner the implementation, without complete details such as the algorithms and their hyper parameters, certain thresholds and the motivation behind them and in some cases, no implementation details are given, but rather an abstract summary of the approach. Moreover, the results claimed lack an evaluation methodology and a concrete dataset, considering only the outcome and focusing mostly on the numbers.

The purpose of this research is to create a benchmarking methodology for the existing solutions and an objective comparison between them. Therefore, the work aims to report missing key methods and propose alternatives that can conduct to similar results and build a foundation for the future research.

2 State of the Art

Here you will build a relatively complete list of papers that already solve the problem you are trying to solve (or a very similar one). If you think that your problem is very novel (i.e., nobody has attacked it before) please think twice. Most probably you are missing something. However, if indeed your problem has never been tackled before, you have to convince the reader how come this is the case. A good state-of-the-art section lists papers, giving a short summary of the contributions of those papers are (i.e., what they offer to the literature). After you list those contributions, you have to say how they fall short (i.e., why they do not solve your problem) and what you think your thesis will do to tackle your problem.

The starting papers for my thesis and the ones that will be analysed and criticised are presented next, together with a short summary of their contribution and why I consider they need improvements.

The Cupid framework [**madhavan2001generic**] has a very interesting approach because it starts from the taxonomy introduced by Rahm and Bernstein [**rahm2001matching**] and evaluates the framework based on the comparison with other frameworks that were introduced until that time. Their approach is focused on schema-based elements and they introduce multiple methods such as weighted similarities based on linguistic and structural matching and schema trees converted in Direct Acyclic Graphs (DAG). The paper tries to detail the methodology, but key factors such as multiple thresholds used, how they connect certain methods and the choices in terms of algorithms and thesaurus are missing.

3 Thesis Objective

The objective of the thesis is to make a straight comparison between different schema matching techniques and to build a foundation for the future research. Through this research, I will help disambiguate the schema matching methodologies, while building an effective framework to compare them. More specifically, my contribution is the following:

- An extensive literature survey
  Starting from the taxonomy build by Rahm and Bernstein [**rahm2001survey**], I will extend the taxonomy and improve it by adding and highlighting the new schema matching methodologies.
- A concrete implementation of selected methodologies
  The implementation purpose is to disambiguate the methods by adding the missing details and creating a complete solution. Moreover, the implementation will be publicly available in the research community.
- A benchmark dataset which will be used to evaluate the methodologies
  An important detail that is generally missing from the previous research is the dataset. Therefore, I will create a public dataset that can be used for all the algorithms implemented and I will detail its advantages and limitations.
- An evaluation approach
  I will design a benchmarking framework that will be used to evaluate all the methodologies.
- Present the results
  The methods will be evaluated according to the designed approach on the benchmark dataset and the results will be objectively analysed.

4 Milestones and Deliverables

How are you going to approach your thesis project? Make a concrete plan and insert a Gantt chart describing the various (high level) tasks that you plan to tackle your research problem and write your thesis.