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

An objective evaluation of schema matching techniques
Thesis Proposal

2019-04-29

1 Introduction

Nowadays, it is becoming more and more common to encounter enterprise settings such as data warehouse, where there are hundreds of different databases. Thus, enabling a seamless integration of the data assets has become the goal in database research area such as data integration. Traditionally, data integration used to be performed manually, but now giving the tremendous number of entities, the task results in an awfully labour intensive effort.

Therefore, schema matching has become the main component in data integration, because it relates to identifying corresponding elements and relationships in different schemas. Currently, multiple algorithms and systems have been developed to address this issue, from semantic matching to clustering and machine learning algorithms.

However, the complexity of the matter relies on the ability to distinguish a match from a non match. Commonly, the match operation is analysed syntactically, but the semantics perform also a crucial role in the process of entity disambiguation. Moreover, some methodologies are using either the syntax or the semantics to find similarities, while others have been focusing on using the unsupervised learning techniques, such as clustering. Besides, in order to find the true matches, the early research has employed the traditional machine learning techniques as well as the popular similarity measures.

Nevertheless, the existing solutions approach the problem of schema matching 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. In some cases, the implementation details are completely missing and a rather abstract summary of the approach is given. 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

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 [4] starts from the taxonomy introduced by Rahm and Bernstein [7]. It describes a very interesting approach and evaluates the framework based on the comparison with other frameworks that were introduced until that time. Their methodology 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 slightly 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.

In 2005, Madhavan et al [5] proposed a methodology where they used existing schema information to enhance the matching for new schema elements. Their approach is schema-based and instance-based and it uses machine learning algorithms to determine the similarities between elements. Their methodology is based on the methodology proposed in 2001 by Doan et al [2]. In the paper, they describe the algorithms employed together with their limitations and they propose a methodology for testing the framework. Considering the papers as two completely independent research studies, highlights the lack of complete information regarding certain aspects of the research or testing methodology. However, the papers complement each other and together they reflect an entire overview of semantic based schema matching. My goal is to replicate the research based on the two papers, enhance it with the current algorithms for semantic matching and highlight

the discrepancies.

From supervised learning to unsupervised learning, Zhang et al [9] researched a new approach for schema matching based on clustering. They focused not only on the schema elements, but on the schema relations as well. They consider that the relations between tables can help understanding if two columns are a true match or a false positive. They partition the data into distribution clusters based on the distribution similarity metric introduced in Zhang et al [8], called Earth Mover's Distance (EMD). It is a complex research, based solely on mathematics that can represent a challenge from the perspective of computer science. In the evaluation section, they indicate the datasets and their sources and specify the metrics used (precision and recall). An interesting aspect in the evaluation section is the fact that they computed several views from the data, in order to enrich the datasets which can add potential bias to the results.

In 2015, Microsoft [1] published a research that claims to help enriching the databases by finding a match between their schemas and some spreadsheets. The spreadsheets contain potential information that can help to disambiguate the schema elements. They create a new similarity metric based on Jaccard Similarity, test it in their own environment and report good results. The challenge is finding if the approach is a good solution in a different environment or if the similarity metric can perform good in other settings.

In 2018, Fernandez et al [3] construct a semantic matcher by using word embedding techniques combined together to find semantically closed word groups. The research is focused on matching multiple words together, thus they discard the typical word embeddings technique. The decision originates from the claim that the method works only for single words. Therefore, they create "coherent groups" which represent the technique of combining word embeddings that proved to be working in practice. One example is the cosine similarity, used to compute the "coherent similarity" between a set of vectors. Moreover, they use other two matching techniques: an instance based matcher that uses Jaccard similarity and a syntactic matcher. One interesting aspect that I have noticed in the paper is the fact that they claim to use a state-of-the-art method, but they refer to an entire survey. Moreover, alike the previous papers, they describe formulas, but they do not specify threshold values or their limitations. Their evaluation section describes the datasets used together with the source and briefly mention the methodology, focusing on the results.

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 [6], 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 researches 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

To successfully achieve the goals proposed, I plan to dedicate one month for building a new taxonomy, one month to implement each methodology, one month to design the evaluation framework and one month to evaluate and report the results. The planning can be viewed in the Gantt chart below:

	Start Date	End Date	Timeline	Status
Thesis breakdown	May 1, 2019	Dec 1, 2019		
Implement semantic method	May 1, 2019	May 31, 2019		Upcoming
Implement syntactic method	Jun 1, 2019	Jun 30, 2019		Upcoming
Implement Cupid methodology	Jul 1, 2019	Jul 31, 2019		Upcoming
Implement clustering	Aug 1, 2019	Aug 31, 2019		Upcoming
Build taxonomy	Sep 1, 2019	Sep 30, 2019		Upcoming
Build evaluation methodology	Oct 1, 2019	Oct 31, 2019		Upcoming
Evaluate and report results	Nov 1, 2019	Nov 30, 2019		Upcoming

Abbildung 1: Gantt chart with high-level tasks

Literatur

[1] Eli Cortez u. a. „Annotating database schemas to help enterprise search“. In: *Proceedings of the VLDB Endowment* 8.12 (2015), S. 1936–1939.

[2] AnHai Doan, Pedro Domingos und Alon Y Halevy. „Reconciling schemas of disparate data sources: A machine-learning approach“. In: *ACM Sigmod Record*. Bd. 30. 2. ACM. 2001, S. 509–520.

[3] Raul Castro Fernandez u. a. „Seeping semantics: Linking datasets using word embeddings for data discovery“. In: *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE. 2018, S. 989–1000.

[4] Jayant Madhavan, Philip A Bernstein und Erhard Rahm. „Generic schema matching with cupid“. In: *vldb*. Bd. 1. 2001. 2001, S. 49–58.

[5] Jayant Madhavan u. a. „Corpus-based schema matching“. In: *21st International Conference on Data Engineering (ICDE'05)*. IEEE. 2005, S. 57–68.

[6] Erhard Rahm und Philip A Bernstein. „A survey of approaches to automatic schema matching“. In: *the VLDB Journal* 10.4 (2001), S. 334–350.

[7] Erhard Rahm und Philip A Bernstein. „On matching schemas automatically“. In: *VLDB journal* 10.4 (2001), S. 334–350.

[8] Meihui Zhang u. a. „On multi-column foreign key discovery“. In: *Proceedings of the VLDB Endowment* 3.1-2 (2010), S. 805–814.

[9] Meihui Zhang u. a. „Automatic discovery of attributes in relational databases“. In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*. ACM. 2011, S. 109–120.