A Holistic Approach to Schema Matching

Bhavik Doshi

Committee Chair: Dr. Rajendra K. Raj
Reader: Dr. Carol Romanowski
Department of Computer Science
B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York

April 29, 2009
Abstract

Schema matching plays an important role in the architecture of data integration and is the process of identifying semantically related objects. It can be described as a method in which source schema elements are mapped with the target schema matching elements. It also plays a critical role in enterprise information integration and has been a popular data management research topic, particularly in building data warehouses and marts. Due to the subjective nature of schema matching, automating the process becomes complex to achieve though efforts have been made to make it semi-automatic. In addition traditional techniques take advantage of only one of the aspects of schema structure, syntax, semantics or data and their probability distribution. By exploiting individual features it becomes difficult to increase the success rate, as each approach is implemented independently of each other.

The latest development in this field is the use of a holistic approach for schema matching which is domain independent and works on the principle of integrating different match processes. Preliminary results suggest that such an approach would outperform traditional algorithms if more research is put on selection of datasets and metrics. The main proposition of this project is to extend the already implemented algorithm and test its feasibility in the real-world scenario. This project would also evaluate the compatibility of the holistic approach with oracle databases to make it more generic and domain independent. This project will compare the performance of the holistic approach with traditional ones in terms of precision, scalability and accuracy.
1 Introduction

The complex nature of schema matching makes it one of the most important steps in the process of data integration. In addition to this schema matching also plays a decisive role in enterprise information integration making it a prime topic in data management research [11]. Schema matching can be defined as, given source and target schemas, matching maps source elements to target elements [12]. Automation of such a process becomes a difficult task to achieve due to its subjective nature of schema matching but constant efforts are being put to make it semi-automatic. Traditional schema matching techniques have exploited individual features of schema structure, syntax, semantics and the data and probability distributions. Exploiting individual features lead to reduction in the overall impact to the process of schema matching and decrease its efficiency.

Sinha et al.[12] propose a holistic domain independent approach to schema matching, which generalizes the process by integrating different match processes. The holistic approach uses two methods one from Kang et al.[6] which is an instance based approach and other from Li et al.[7] which is an element based approach. The instance based approach proposes a two step technique and uses data’s mutual information and probability distribution to perform matching in spite of the presence of opaque column names and data values. On the other hand the element based approach proposes a matching tool which automates the schema matching between the source and target schemas and is based on structures, constraints and other elements.

The holistic approach is based on combination of both instance and element level approaches and develops a generic integrated technique which is domain independent for relational databases. This approach uses techniques from both the above methods along with semantics and structural aspects of relational databases to develop a holistic approach. With databases poorly built in the real world it becomes difficult to do schema matching but the holistic approach shows better results in spite of relational design, data values and database object names. Initial results show that the holistic approach generates better results than the above mentioned individual techniques. Precision and recall were the metrics used and java scripts were written to test the accuracy of the algorithms.

![Figure 1: The Holistic Algorithm Approach](image-url)
The core of the project is to use the accomplished work on the holistic approach and extend it further to customize the schema matching process and also to visualize and interact with large real-world relational databases comprehensively. This is because preliminary results suggest that such a matching process would outperform traditional approaches if appropriate datasets are brought into picture. For the same purpose, relational datasets with sufficient number of attributes and tuples will be used to test the accuracy and scalability of the holistic approach. Also conversion of non-relational datasets to relational datasets would be done to evaluate the matching process. Once the datasets are ready, they will be tested with all the approaches and performance characteristics of all the processes will be measured. As stated before, precision and recall will be used as the base metrics to measure the correctness of the holistic approach. All the test results and evaluations will be then reflected as a chart in the final project report.

2 Related Work

Schema matching has become an integral part of data management research and considerable amount of work has been done in the same. Apart from the instance based approach by Kang [6] and elemental based approach by Li [7], several different matching algorithms have been developed. Melnik et al.[9] propose a graph matching algorithm in which graphs are created from constraint and schema information. The major advantage of such an approach is that it can handle non-relational datasets unlike most of the approaches. Rahm et al.[11] describe a complete taxonomy of various schema matching approaches and under which the holistic approach can be best described as the hybrid approach.

In their paper Li and Clifton [8] describe the use of neural networks to find semantic similarity between elements but on the downside this technique requires apt cleaning and training the dataset model before use. The algorithm developed by Palopoli et al.[10] describes an interactive framework which creates a global integrated abstract schema. The specialty of such a framework is that it can extract mappings from multiple data sources and not just the source and target schemas. On the other hand this technique requires setting up of initial basic rules manually for the algorithm to process. One more element based approach apart from Li [7] is by Bohannon et al.[1] which makes use of logical annotations for schema matching. They call such a type of matching as contextual schema matching but when it comes to weakly named schemas, this technique fails to handle then using logical annotations.

Another Holistic approach is proposed by He and Chang et al.[3] in which they develop a framework for matching query interfaces on the deep web. For the same purpose they have developed a different technique [4] which uses data mining techniques like clustering to identify a match between similar columns. In such a framework the choices of distance metric determine the quality of the result if the data quality is poor. Drumm et al.[2] propose the use of ontologies to develop a QuicMig technique for schema matching. Using ontologies and thesaurus can improve the match quality to a considerable extent but at the same time it is a tedious process and requires expertise and a lot of domain knowledge.
The holistic approach proposed by Sinha et al. [12] combines the instance and elemental based approaches and develops a generic, integrated and a domain independent schema matching technique. The diagram below depicts the architecture of the holistic approach and it is also supported by a graphical user interface. The mentioned framework used relational datasets and tests the accuracy of the framework with the help of precision and recall metrics. The holistic approach fills gaps of both instance and elemental based approaches and gives a combined unified approach to schema matching. I plan to use the mentioned architecture in [12] and propose to test the schema matching algorithm for larger real-world relational datasets.

![Figure 2: The Holistic Approach Architecture [12]](image)

3 Hypothesis

From the analysis made in [12] it seems there is a need to determine on whether or not the holistic approach outperforms instance and elemental based approaches, when it comes to large real world relational datasets. The diagram above depicts the basic architecture of the holistic approach which takes into account several aspects of relational design and database objects in order to obtain an integrated schema matching.

The holistic algorithm makes a start with the instance based approach [6] and limits the number of connections in the dependency graph. A dependency graph for a particular schema can be defined as the complete graph consisting of calculated mutual information between each pair of nodes. Mutual information can be termed as a constructive metric as it represents all the information a node knows about the other, but on the other hand calculating for all pairs would slow down the algorithm drastically. Here the holistic approach would just draw an edge if either the nodes belong to the same relation or they share a referential integrity constraint. Thus this reduces the computation of the algorithm to a considerable extent and in addition also allows the structural properties and relationships of the schema to contribute to the final matching [12].

In the next step the holistic approach revisits the elemental based approach by removing the star schema restriction thus setting the relation in the schema graph to the highest degree as the fact
table. It also restricts the matching of similar data types by assigning values ranging from 0 to 1 depending on the extent of matching. Thus the holistic approach uses both the above approaches with many other optimizations to improve the extent and quality of schema matching. With all this the holistic approach also takes into consideration semantic similarity and constraint based methods which contribute to the matching in the form of weighted average.

On the other hand the holistic approach is sensitive to many factors like data distribution, amount of data and database object names. Hence conclusive tests have to made on such an approach and then compared with traditional approaches. The metrics which will be used during the course of the project will be precision and recall. Precision can be defined as the fraction of correct matches to the total number of generated matches by the holistic approach. On the other hand, recall is the fraction of correct matches to the real number of correct matches. These metrics were used by Kang et al. in their evaluation of the instance based approach.

All the three algorithms mainly depend on the two decisive factors when it comes to relational datasets namely, the number of tuples in the relations and also the number of relations in the schemas. Hence it becomes of utmost importance to vary these factors and then analyze the performance and accuracy of the three algorithms. Scaling the number of tuples in the relational datasets showed that the holistic approach showed better results when it came to recall but not in precision. The main reason here is that the amount of false positive matches generated by the holistic approach were high. It is likely that if naming conventions and relational design are altered to suit the requirements of the holistic approach, then it would outperform other approaches in precision. All these factors show the need to evaluate the holistic approach with large and different types of relational datasets so as to understand the correctness of the algorithm in the real-world scenario.

In addition to this, another scenario can be to evaluate the holistic approach by scaling the number of relations in the schemas or when the attributes and data values are completely opaque. Preliminary results suggest that the holistic approach would suit a better replacement to traditional approaches but it would require careful and through investigation with real world relational datasets as such an approach is sensitive to many relational data factors. The tables below compare all the three approaches using both, precision and recall metrics.

![Figure 3: Precision results](image)
4 Solution Design and Implementation

Figures 3 and 4 show that the holistic approach would surpass both the instance and elemental approaches if a systematic approach is followed while evaluating it. Hence there is a need to identify the sensitivity of the holistic approach and evaluate it with a series of different real-world relational datasets. As already stated, although theoretical analysis show that the holistic approach is better [12], no real-world scenarios with datasets have been tested against the implemented system.

During the course of the project work, I will use many relational datasets to conduct experiments and I might also have to convert non-relational datasets to relational ones. The project has reached to a stage where I can mention some of the datasets which will be used during the evaluation process. For instance, I would use KDD cup datasets from University of California, Irvine [5]. Some of the datasets mentioned here are non-relational datasets and in the process I might have to convert them into relational ones. Some of the examples of relational datasets mentioned in [5] are E Coli Genes, M Tuberculosis Genes and Movies.

4.1 Testing and Evaluation

Once modifications have been made on the datasets, tests would be conducted on all the three approaches. As already mentioned, partial experiments on the holistic approach with fairly small number of datasets have been made by [12], but it does not seem to comprehend the success of the holistic approach. The goal of this project would be to assemble considerable amount of relational datasets with varied number of rows and tuples and then test it against the three approaches to give clear explanations on the feasibility and efficiency of the holistic approach. The three major characteristics which would be taken into account are defined below:

- Accuracy in terms of generating the correctness of the matching’s for all the three approaches. This becomes extremely important as several different possible combinations of source and target schemes would be used.

- Scalability in terms of varying the size of data and the number of relations present between them. It is important to verify the modifications required in the holistic approach when one changes the above parameters so as to test the feasibility of the holistic approach.
Performance in terms of calculating the computational complexity of the holistic approach compared to other approaches.

Finally, to comprehend the goal and provide a detailed comparison of the Holistic approach with the instance and elemental based approaches, the algorithm would be tested for its efficiency and correctness with the help of two metrics, precision and recall. All the above evaluation criteria would be applied to the holistic approach to test its accuracy and scalability compared to the traditional approaches.

5 Roadmap

Below specified is the list of deliverables for the project:

- A technical report consisting of the list of datasets used to evaluate the holistic approach, evaluation procedure and comparative results derived from the above evaluation.
- A fully-functional code of the already implemented algorithm by [12] and modifications made on the code to suit the real-world datasets.
- A complete chart of results derived and full comparative results for the three approaches.
- A final PowerPoint presentation giving the overview of the project.

Consequently, I will regularly update my website with all the developments made in the project.
5.1 Project Schedule

The current development and future plans of the project are shown in Table 1.

<table>
<thead>
<tr>
<th>Time Frame</th>
<th>Modules</th>
</tr>
</thead>
<tbody>
<tr>
<td>February 2009</td>
<td>Project Preproposal</td>
</tr>
<tr>
<td>March 2009 - May</td>
<td>Obtaining the pool of Relational datasets</td>
</tr>
<tr>
<td>May 2009</td>
<td>Project Proposal</td>
</tr>
<tr>
<td>May 2009 - July</td>
<td>Finalizing the set of relational datasets from the pool</td>
</tr>
<tr>
<td></td>
<td>Customized Hollistic approach design</td>
</tr>
<tr>
<td></td>
<td>Improved code for the Hollistic approach</td>
</tr>
<tr>
<td></td>
<td>Code for converting Non-relational datasets to Relational ones</td>
</tr>
<tr>
<td>June 2009 - August 2009</td>
<td>Final Testing and Analysis</td>
</tr>
<tr>
<td></td>
<td>Evaluating the Performance of the Holistic Approach</td>
</tr>
<tr>
<td></td>
<td>Detailed description of experimental results by the means of graphs</td>
</tr>
<tr>
<td>August 2009</td>
<td>Project Report</td>
</tr>
<tr>
<td></td>
<td>Defence</td>
</tr>
</tbody>
</table>

References


