A COMBINED MAPPING APPROACH FOR RELATIONAL SCHEMAS

M.Y. Chaudhari and V.M. Shelake
Department of Computer Engineering, SES, Group of Institutions Faculty of Engineering Diksal, Karjat, Raigad, Maharashtra, India
minalc.1477@gmail.com, vijay_sakec@yahoo.co.in

ABSTRACT

Nowadays many issues of data handling, large amount of data is generated on every day and it is difficult to fit in frame for data mapping. Still there are inconsistencies and incompleteness issues in schema mapping. This paper presents probabilistic approach in the issue of schema mapping. Our approach is indicative, mountable, and extensile. We introduce Combined Mapping Approach for Relational Schemas to this problem using state of-the-art probabilistic reasoning techniques. We determine that the correctness of Combined Mapping Approach is greater than 33% above of metadata-only method for small data examples, and this method normally finds perfect mappings even if a sector of the data is unpredictable.

Keywords: *Attribute Mapping, Probabilistic Approach, Integration, Data Mapping*

Introduction

Data mapping is the technique of identifying objects which are related to each other. In other words, database mapping is a method of finding the correspondences between the concepts of different distributed, heterogeneous data sources. A database mapper is a binary operation which takes two schemas as input and determines their corresponding elements. Generally, the analysis of all matches between elements of two schemas is not possible due to the presence of semantic in input schemas. Therefore, a database mapper should only provide match candidates through a user interface in such a way that users can accept, reject, or change them. The several semi-automatic database mapping algorithms have been proposed to determine matches between different elements of two schemas [1][2][3].

Enterprise data is getting more diverse and at the same time, it is advantageous for businesses to leverage data and transform it into meaningful results. However, enterprises today collect information from many sources referring the same entities. To merge this data and make sense of it, data mapping is used which is the process of establishing relationships between separate database schemas.

For example, Microsoft Dynamics CRM has several data sets which encompass of different items, such as Leads, Opportunities, and Participants. Each of these records has some fields like Name, Account Owner, City, Country and more. The application also has a well-defined schema along with attributes, enumerations, and mapping protocols. Hence, if any record is to be inserted to the schema, a data map wants to be designed from the data source to the CRM account.

Depending on the factors like number, schema, and primary keys and foreign keys of the relational databases records, database mappings can have a changing degree of complexity.

For example, data from three various databases tables are combined and mapped to an Excel destination. Thus, attribute level information as well as same entities are to be attempted for the actual database mapping [4][5].

In simple words, data mapping is a relationship between two or more data systems. Data mapping joins two different types of data models together. In data integration procedure data mapping is one of the most significant factors. Making it simpler, data mapping is all about searching how a database connects to another database. To make it easier and simpler to understand the concept, here is an example.

Suppose you have a list of people, a list that have some names of your friends. As well you have another list that have the phone numbers of your friends. Now you want to combine both of the lists and that is where data mapping works for you.

Data mapping is the initial step in many compound segments to start the integration process. That contains data transformation between the data source and the data

destination. It helps in the detection of a private data such as the last digits in a social security number and so on. It can also be used for securing many databases and combine them into one while looking for redundancy [6].

Literature Survey

To understand the various data mapping techniques, here in review of literature the discussion of some schema matching techniques are necessary. The various mapping techniques like complex schema matching, Entity matching, mapping between different elements of two schemas, record matching, probabilistic approach of schema mapping, Automatic schema matching are discussed below.

Review of Literature

The various important approaches of data mapping are discussed as follows:

Complex Schema Matching i.e. CSM, the problem of searching semantic correspondences between elements of two compound schemas, plays important role in many applications, such as data warehouse, heterogeneous data sources combination and semantic web. The existing way to automating schema matching focus on calculating direct element matches between two schemas.

However, this work involves the relationships between real-world schemas involving many complexes matches besides 1:1 matches. The limitation still remains on matching efficiency, because the candidate matches space is so large which they need searching [1].

Entity matching as well as value mapping across two varied data sources are critical tasks in applications involving data combination, data warehousing, and federation of databases. Before data can be combined from multiple tables, the columns and the values appearing in the tables must be matched. Anuj Jaiswal, David J. Miller anticipated a novel method that optimizes embedded value mappings to improve entity matching in the occurrence of opaque data values and column names. In this approach, the fitness objective for matching a pair of attributes from two entities depends on the value mapping function for each of the two attributes [2].

The work includes schema matching, which provides a mapping between different elements of two schemas. It is complicated due to the existence of semantic in input schemas. It is achieved through different matchers for merging databases. It determines all matches between elements of two schemas. Still, it there is scope for improving accuracy [3].

First schema matching and Second record matching are two essential steps in integrating many relational tables of different schemas, where it first unifies the schemas and then it detects records mentioning to the same real entity. The two processes have been thoroughly studied separately, but in this work, they had done combination of both. They have used indexing and estimated the matching likelihood between two records [4].

When link the data stored in varied database, very difficult problem is entity matching, i.e., matching records representing semantically corresponding entities in the real world, across the sources. When decision tree techniques have been used to understand entity matching protocols, most decision tree learners have an inherent representational bias, that is, they generate unvaried trees and prevent the decision boundaries to be axis-orthogonal hyper-planes. Cascading other classification approaches with decision tree learners can improve this bias and potentially rise classification accuracy. In this method user apply a recently developed constrained cascade simplification method in entity matching and report on observed evaluation using real data. Its results show that this method smartly performs the base classification methods in terms of accuracy, especially in the dirtiest case [5].

The probabilistic approach of schema mapping is declarative, scalable, and extensible. This approach designs upon recent results in both schema mapping and probabilistic reasoning techniques in both fields. The problem of choosing the best mapping from a apart from potential mappings, given both metadata constraints as well as data samples. As selection has to reason about the inputs and the dependencies between the chosen mappings a new schema mapping optimization problem which gets interactions between mappings then they familiarize Collective Mapping Discovery (CMD), solution get using probabilistic

reasoning techniques, which allows for discrepancies and incompleteness [6].

Automatic schema matching concept is typically only deals with result in attribute correspondences. However, real world data combination problems often require matching whose arguments want all three types of fundamentals in relational databases and that are relation, attribute and data value. In this concept the definitions and semantics of three extra correspondence types regarding both schema and data values. Two methods for automatically recognizing these correspondences are developed. First requires some limited number of duplicates across data sources. Another is a general instance-based method with no such requirement [7].

Proposed System

As we see that many data mapping methods use to analysis or simplify the complex data. Day by day the data is generated in large amount and it is very difficult to analysis so to simplify mapping of data here data type & attribute are used very effectively.

Problem Definition

In this proposed system, combined database mapping approach for interoperating among different database systems. In this proposed system, the database mapping process uses of three main phases: -attribute mapping, data type mapping and data merging.

The attribute mapping determines whether the names of two attributes are similar or not by calculating the percentage of similarities between their names. It uses both structural and semantic level information including string mapping, dictionaries to calculate the percentage of similarities between the names of two attributes. Then, our approach compares the similar mapping attributes and data type mapping and selects the better match. The final phase of the attribute mapping process is merging, which merges those pairs of attributes that are identified matched, while those attributes that are not matched remain as they are.

System Architecture

- **User**

The users basically include engineers, scientist, business analytics and others who thoroughly familiarize themselves with the facilities of the concept in order to implement their application to meet their complex requirement. These users try to learn most of the Data mapping facilities in order to achieve their complex requirements.

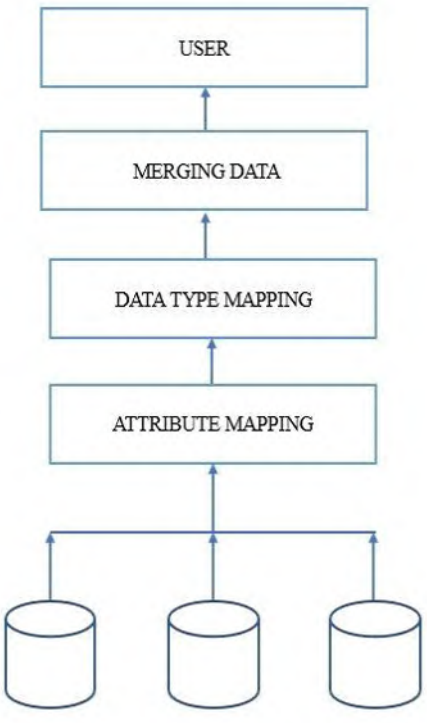

Fig. 1 - System Architecture

- **Merging Data**

In data merging, first we select the two data set (table) to create a single data set for easy reporting & analysis. The first data set that contains some attributes like account number, name and ownership information. The second data set contains attributes like number of discharges and payment details etc. Unfortunately, we do not have a common ID to merge the two data set so we try to merge the data using data frames, as the columns have different names; we need to declare which columns to match for the left and right Data Frames.

- **Steps of Data Mapping**

The algorithm for combined database mapping approach is defined below:

1. Start
2. First select the data bases.
3. Display of attribute & data to be map for respective databases.
4. Select appropriate Attribute from respective databases for further Mapping.

5. Apply Mapping Probabilistic strategy for selected Data databases.
6. Display Data Mapped for selected Data databases.
7. Display of data as perscore of mapping in ascending order.
8. Stop.

Discussion & Analysis

Analysis

Probabilistic matcher determines the best match for each combination. For hospitals data set we are analyzing approximately 7.5 thousand combinations. On the system, this takes about 2 min and 11 seconds to run. The matching is done for three scale best match, worst match &average match < 80 in two data set.

Observations

i)Top Five Best Match

The Fig.2 shows the top five best match found in the two-hospital datasets use in the data mapping. In the graph x-axis represent the data mapping score & y-axis represent the top five individual record.

Fig. 2 - Best Matchin Database

ii) Top Five Average Match

The Fig.3 shows the top five Average match found in the two hospital datasets use in the datamapping. In the graph x-axis represent the datamapping score & y-axis represent the top five average case individual record.

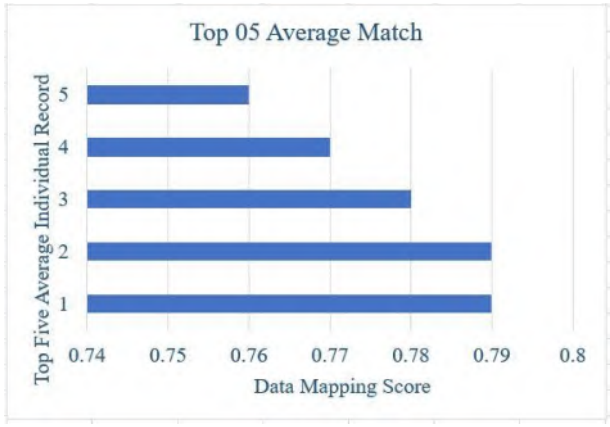

Fig. 3 - Average Match in Database

The Fig.4 shows the data matching with various scales like best match, average match and worst match.

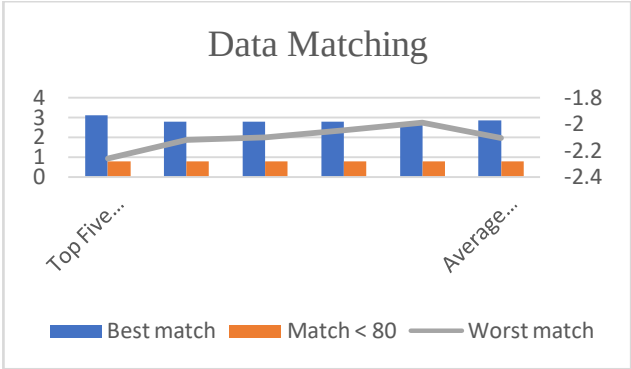

Fig. 4 - Various Matching Scale

Conclusion

Data mapping is always resource-intensive requiring hands-on development, review, and knowledge about all sources and targets. Human intervention is required for mapping design and validation of map results. Marketable and open-source mapping tools can assist in the process by providing changing degrees of automation. Manual review is required, to a varying extent, to map the portions that failed automated mapping and to authenticate the results of automated mapping. Linking various record sets on text fields like names and addresses is a common but interesting data problem. In this work, we had used combined mapping approach for merging databases schemas. Our proposed approach contains structural and semantic level information to map database schemas for getting better results compared to existing approaches.

References

1. **Y.Qian ,Y.Li,J.Song, & L.Yue(2009)**, "Discovering complex semantic matches between database schemas," in Proceedings of International Conference on Web Information Systems and Mining, 2009 pages -756-760.
2. **Anuj Jaiswal, David J. Miller, Member (2010)**, "Uninterpreted Schema Matching with Embedded Value Mapping under Opaque Column Names and Data Values", IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 2, 2010 pages 291-309.
3. **Hamidah Ibrahim, Yaser Mohammad Karasneh, Meghdad Mirabi (2014)**"AnAutomatic Domain Independent Schema Matching in Integrating Schemas of Heterogeneous Relational Databases", Journal of Information Science and Engineering2014 pages -1505-1536.
4. **Binbin Gu , Zhixu Li, Xiangliang Zhang, An Liu, GuanfengLiu, Kia Zheng, Lei Zhao, Xiaofang Zhou(2017)** "The Interaction between Schema Matching and Record Matching in Data Integration", IEEE Transactions on Knowledge and Data Engineering ,2017pages -186-199 .
5. **Khai Nguyen and Ryutaro Ichise(2016)** "ScLink: supervised instance matching system for heterogeneous repositories," National Institute of Informatics, Vol. 32, 2016, pages -830-863.
6. **Angelika Kimmig, Alex Memory , Ren´ee J. Miller , Lise Getoor(2017)**" A Collective, Probabilistic Approach to Schema Mapping" in IEEE 33rd International Conference on Data Engineering2017, pages 921-932.
7. **Aibo Tian, Mayank Kejriwal, Daniel P. Miranker(2014)** "Schema Matching over Relations, Attributes, and Data Values"In conference on Scientific and Statistical Database Management, SSDBM '14, Aalborg, Denmark, June 30 - July 02, 2014, pages 28:1–28:12.