Automatic Entity Mapping for Scalable Data Exchange

K. Gowthami
M.Tech
Computer Science and Engineering
JNTU College of Engineering

K. Madhavi
Associate Professor
Computer Science and Engineering
JNTU College of Engineering

Abstract Data Exchange creates an instance of a target schema from an instance of a source such that source data is reflected in the target instance. The approach for data transformation is based on mapping techniques and using schema mapping expressions representing high level relations between source and target. Data exchange is the process of taking data structured under a source schema, and generating an instance that adheres to the structure of a target schema. The prevailing approach for this process is based on schema mappings – high level specifications describing relationships between Source and Target areas. The mapping techniques unable to resolve ambiguous exchange scenarios and entity fragmentation .To address this problem, the proposed method based on mappings and privacy preserving data exchange that employs the best relations that can host source instances. The mapping method out forms other methods in terms of quality and scalability of data exchange and it improves the privacy during data exchange scenarios.

Index Terms—Data exchange, schema mapping, scalability, tree representation

1. INTRODUCTION

Now a day's the continuous growth in data demands the need for the integration of structured data with the goal of making data that available from various independent and heterogeneous resources [1]. Data mining is a multidisciplinary subfield of computer technology. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database system. The overall goal of the data mining process is to retrieve data sets and exchange it into an understandable structure for the further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, virtualization, and online updating. Data mining is the analysis step of the "Knowledge discovery in databases" process, or KDD. Data are any facts, numbers, or text that can be processed by a computer. Today, businesses are accumulating extensive and growing amounts of information in unique codecs and different databases. This includes: Operational or transactional data such as, sales, cost, inventory, payroll, and accounting. Nonoperational data, such as industry sales, forecast data, and macro economic data. Meta data - data about the data itself, such as logical database design or data dictionary definitions The patterns, associations, or relationships among all this data can provide information. As an example, evaluation of retail factor of sale transaction facts can yield data on which products are promoting and when facts can be converted into knowledge about historic styles and destiny tendencies. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts. The data integration through exchange methods which creates an instance in both source and target schemas. The data exchange problem has also been studied by Fagin etal., who propose theoretical foundations behind data exchange [4].The target solution generated by a system is pruned and Processed to generate the core solution through introduced the concept of universal solution in the post-processing approach to compute the core solution [6], [7].As argued in [8], this technique may impairs efficiency of a data exchange system. On the other hand, in preprocessing approaches such as ++Spicy [8], schema mapping expressions are directly generate the core solution and it refined the mappings. In spite of considerable Improvements in data exchange, there exist scenarios that cannot be handled properly. More specifically, in many existing data exchange systems based on schema mapping (e.g., [2]), first mappings are created based on schema level information, and then these mappings are used to translate source to the target data. In this approach, schema mappings are treated as independent expressions and many mappings overlap each other. Data mappings are implemented by using mapping tables in which
attributes being mapped. The tuples in the mapping tables show values in the mapped relations and the values are correspondence to each other. These tables are treated as constraints (aka mapping constraints) on the exchange of data between peers. In [9], a sample-driven schema mapping technique is proposed that builds the schema mappings from sample target instances. This technique generates mapping expressions given each pair of sample source and target instances. In EIRENE [10], data examples are used to refine schema mappings. The main difficulty when attempting to characterize data transformation using data examples is likely to describe the same behavior. The existing uncorrelated mappings that may result in duplication of data and loss of associations in data exchange. Map Merge[11] exploits constraints in the source and target schemas to find the associations and improves the quality of mappings and increases the scalability. In generalization relation null values may occur if it is realized through materializing all specific classes inside a single table and which leads to ambiguous mappings and incorrect data exchange. When exchanging incomplete data and mapping inversion in the source may arise null values [14].The problem of entity fragmentation, and the inability to resolve ambiguous data exchange scenarios caused by different implementations of a generalization relation in source and target, are consequences of ignoring data level mappings. The gap between data level and schema level mappings in schema mapping-based data exchange results in semantic heterogeneities, and consequently, incorrect and redundant target instances. The Scalable Entity Preserving Data Exchange method we propose bridges the gap between data level and Schema level approaches in data exchange. This system avoids entity fragmentation in data translation, and resolves ambiguous data exchange scenarios, which are consequences of different implementation of generalization relations in the source and the target. Finally it report an extensive set of experiments to show how entity preserving data exchange out forms existing works Frequent Itemset mining (FIM) is a core problem in association rule mining (ARM), sequence mining etc., the high calculation and input/output intensity is needed in processing of FIM fastly, is not easy because the FIM consumes the particular unit of time for mining. The data in data miming databases are going to be increasing gradually, and the mostly used

sequential FIM algorithm for processing of data executes on a one machine, these leads to the problem, that suffer from lagging in execution delay. So, by assigning a huge dataset over the cluster it will capable of load across all cluster nodes, and also there will be an immense improvement in the execution of parallel FIM. Frequent item set mining algorithms can be categorized into two segments, namely Apriori, and FP-growth schemes. Apriori is a simple algorithm using the produce and check process that gives a great number of possible itemset; Apriori has to check item set continuously in an entire database. To lessen the time usage for checking databases, introduced a novel approach called FP-growth, which ignores producing candidate item set. FP-growth is the frequent patterns algorithm which is an efficient algorithm for mining the frequent dataset. These method will process the whole set of required, minimized and important information in a scalable manner. It stores all the type of information using the tree extended based structures. The predominating approach for this process is based on schema mappings, which are high level expressions that describe relationships between database schemas. There are two problems while performing the data exchange using scheme mappings: (1) Unclear data that is ambiguous data like zero or missing values, in which properties acquired from inheritance using unrelated associations result from using several approaches and (2) fragmented entities, represents that the data about the single object is discussed over the several tuples in the destination. Current re-clustering techniques totally avoid the data compactness in the region between the small grouped clusters (grid cells) and might combine those small grouped clusters (cells) which are near to each in correspondence with other, but at that same time only they are segregated by less region of small compactness. To issue this problem, Chen introduced an enlargement to the grid-based D-Stream algorithm based on the area of interest between side by side grid cells and shows it's consistence. Structure between two micro clusters have been caught externally by grouping the similar data together using the clustering concept and the results formed from this technique which is helpful for doing the re-clustering concept for micro clusters. Using a shared-density based re-clustering approach is supposed to be the best approach for data stream clustering in data exchange processing. Here the retrieved information is shown in the form of a graph, the total data and the performance calculated has been shown by using the bar charts. The graph gives the entire information about the data exchange scenarios that are relevant on the basis of the frequent dataset mining.

II RELATED WORK

Alexe,B.TenCate,B.Kolaitis,P.G. Tan, [1]proposes a System that supports a methodology for designing schema mappings that departs significantly from the methodology used by existing systems that are described above. In Eirene, a schema mapping is derived from data examples that are provided by the mapping designer. One of the relevant parts of the device is a module that, given a fixed of facts examples, either returns a "satisfactory" fitting schema mapping, or reports that no fitting schema mapping exists. Alexe, B., Hernndez, M., Popa, L., Tan, W. C.: Map Merge[2]incorporates a phenomenon of integration or exchange of data is to design the mappings that describe the relationships between the source schemas or formats and the desired target schema. The Map Merge, that can be used to correlate multiple, independently designed schema mappings of smaller scope into larger schema mappings. This allows a extra modular construction of complex mappings from diverse styles of smaller mappings inclusive of schema correspondences produced by using a schema matcher or pre-present mappings that had been designed by way of both a human person or through mapping equipment. Arocena, P. C., Glavic, B., Ciucanu, R., and Miller, R. J [3] study the maturity of the data integration field it is surprising that rigorous empirical evaluations of research ideas are so scarce. It identifies a major roadblock for empirical work - the lack of comprehensive metadata generators that can be used to create benchmarks for different integration tasks. This makes it tough to compare integration solutions, apprehend their generality, and apprehend their overall performance. Augsten, N., Bhlen, M., Dyreson, C., and Gamper, J [4] Conjunctive queries and views, and we investigate the problem of query answering using views in the presence of dependencies and in particular the problem of finding equivalent and maximally contained rewritings of a query using a set of views in the presence of dependencies. We present an efficient sound and complete algorithm CoreCoverC which finds equivalent rewritings with the minimum number of sub goals in the presence of weakly acyclic local as view tuple generating dependencies. Barbosa, D., Mendelzon, A. O., Keenleyside, J., and Lyons, K. A To Xgene [13] Synthetic collections of XML

Documents have many applications in benchmarking, testing and evaluating various algorithms, tools and systems. Moreover,

Different applications require different documents, with different complexities, sizes, etc. For instance, a benchmark for data intensive applications might require a large and relatively homogeneous document, with many references among elements,

while an adequate test suite for a parser might be an heterogeneous collection with thousands of documents with varying sizes. Given the complexity of writing and/or customizing hard-coded synthetic data generators for specific scenarios, we believe a declarative tool for generating synthetic XML documents will prove useful. ToXgene is a template-based tool for generating large, consistent synthetic collections of complex XML documents. Arenas, M., Prez, J., Reutter, J., and Riveros, C[6] proposes schema mapping is a specification that describes how data from a source schema is to be mapped to a target schema. Schema mappings have proved to be essential for data interoperability tasks such as data exchange and data integration. The research on this area has mainly focused on performing these tasks. However, as Bernstein pointed out , many information-system problems involve not only the design and integration of complex application artifacts, but also their subsequent manipulation. Driven by this consideration, Bernstein proposed in a general framework for managing schema mappings. Fagin, R., Kolaitis, P. G., Miller, R. J., and Popa, L [7] study that data integration systems provide access to a set of heterogeneous, autonomous data sources through a so-called global schema. There are basically two approaches for designing a data integration system. In the global-as-view approach, one defines the elements of the global schema as views over the sources, whereas in the local-as-view approach, one characterizes the sources as views over the global schema. Fagin, R., Kolaitis, P. G., and Popa, L[8] using an approach information integration is to precisely specify the relationships, called mappings, between schemas. Designing mappings is a time-consuming process. To alleviate this problem, many mapping systems have been developed to assist the design of mappings. Fagin, R., Kolaitis, P. G., and Popa, L[9] consider the following scenario for a mapping system: given a source schema, target schemas, and a set of value correspondences between these two schemas, generate an executable transformation to compute target instances from source instances. Gottlob, G., and Nash, A [10] proposes a Data exchange is concerned with the transfer of data between databases with different schemas, according to declarative specifications known as schema mappings. Unlike virtual data integration, concerned with query translation among distributed databases, data exchange aims at actually materializing a target database, for the later use offline. Describing the state of the art in the location of core computation for statistics exchange. Two main approaches are considered: post-processing core computation, applied to a canonical universal solution constructed by chasing a given schema mapping, and direct core computation, where the mapping is first rewritten in order to create core universal solutions by chasing it. Finding the similar itemset in any databases, algorithm like Apriori algorithm is a best way of finding frequent datasets in a database. A variety of Apriori - like algorithm aims to shorten database scanning time by reducing candidate itemset. In the inverted hashing and pruning algorithm, every k-itemset within each transaction is hashed into a hash table. Berzal et al. designed the tree-based association rule algorithm, which works an effective data-tree structure to store all itemsets and to lessen the time used for checking databases. In proposed process with parallel mapping

schema relations and tuple relations based access the user query.
The Apriori like based algorithms produce a huge number of itemsets, where excessive number of candidate itemset will
create a ambiguous situation of finding similar datasets among them, so to give the best result performance of apriori algorithm. Han et al proposed a novel approach called Fp-growth which eliminates the producing of huge number of candidate itemset by projecting the database into a small data structure and then using divide and conquer method for similar itemset to obtain.
Parallel mining algorithms based on apriori, in which the count distribution on all candidate itemset is to be calculated over
Each processer of parallel system computes the internal supporting counts. Each processor has the responsibility to compute support counts by sending local database partitions to all other processors. To diminish time utilization for examining
Databases and trading candidate itemset, FP-development based parallel calculations were proposed as a replacement to the
Apriori based parallel calculations. This issue winds up noticeably articulated with regards to huge and multidimensional databases. Here we are using large number of huge data is to be used and analyzing performance for given data and also finding frequent item sets. This prompts an issue when the information focuses inside every cell aren't consistently dispersed and two near cells are isolated by a small density. After the obtained values the dataset is to be updated with frequently occurring item sets. We have to find the Schema mapping relations and source in a database is mapped from target schema. Schema mapping includes the invention of query or set of queries that transformed to the source information. A interactive mapping creation paradigm is normally precise with that worth correspondences confirmed how a value of target attribute can also be considered on created values form source attributes

III. PROPOSED ARCHITECTURE

The proposed scheme for data exchange using decision trees which gives the solution based on combining both the data level and schema level information. Here we are using hadoop distributed file system databases for large number of datasets to use, processed over large databases and analyzing performance for given data and also finding frequent datasets. Decision tree algorithm will helps in finding the frequent itemset easily and quickly by applying the scalable association rule mining technique in which there follows a measure called support and confidence which gives the value for frequently occurring dataset. Here our process starts with the selection of dataset. The proposed scheme is detailed with the flow diagram Dataset loading and preprocessing Pre-processing is remove null or unwanted dataset from given data's through mining methodology. Choose the dataset for our process and we choose the accident record dataset. The dataset contains states, year, causes, and number of death accidents happened over some period of time from all the states of India and union territories. First upload the information in the table in database. Then apply the preprocessing technique to remove unwanted data in the dataset.

Analyzing and Partitioning of Data

Data splitting is one of common modules using for large no of data processing techniques. Segregating the data into multiple to retrieve the result as efficiently. Here file or dataset is some random number generation based partitioned. Then we analysis the data about the accident and find the overall death rate. Calculate the average death rate by state wise and find the threshold value for the data. Filter the data above the threshold value. The data in the dataset once splitted based on the fragment count, the number of files to be generated based on it. So that from the partitions of a dataset one can easily understood and retrieve the information easily, Here, after analysing the data and partitions based on the specified categories, the data is loaded into the database for performing the operations like mapping, schema mapping, and calculating average tuple relation

Clustering the Data By Threshold

Further the process is, to find the diabetics patient based on symptoms. We next process data with symptoms based on patient record to find the support and confidence measure. Support Measure is important to find the frequent dataset based on itemset. To check the itemset whether present in the frequent dataset, if present to count the itemset. And also measure the confidence for threshold. After that update the dataset on forming of the results that are based on the supporting measures.

Find Density of Micro-cluster

Clustering is like file fragmenting process with finding frequent data through some algorithms and supervised methods or fuzzy clustering methods to be used. Here discussing as Micro cluster process with to which degree the data is analyzed. In this module find shared density of the each micro cluster. Produce the density of the micro cluster result in graph structure.

Re-cluster using shared density

Re-clustering which is the process of doing the clustering again, after performing the formation of similar or homogenous data together into a one group. The re-clustering method will improve the performance of the dataset greatly when one works on large databases like on HDFS. The data here is compressed to a great extent while compared to the results that are obtained by doing the clustering technique. And the shared density here, using the re-clustering concept defines the values having the areas which are representing the regions with high density. The results that are found using shared density using re-clustering will create a maximized composition of similar data. The maximized dataset formation of clusters, are to be joined which have the properties near to each other and they are separated with low area density We carried out experimental results that the dataset are taken from the accident cause. Where the dataset include each and every individual state information (accidents caused through various vehicles and other causes etc.,) of a overall dataset.
Initially we load the dataset, before that, we apply pre-processing for the data to be view in a structured format and the null values to be removed. The dataset includes the number of value based results of about the average death causes both male and female.

Once uploading dataset about the total cause occurred due to various accidents for both male and female is completed, then we perform the clustering, it shows the similar frequent data of the dataset which we loaded. Then we partition the dataset based upon the fragment count for which the data present in the choose partition, so that large data in the dataset is to be fragmented into specified number of fragments. After applying the mapping and schema mapping, clustering should be done to calculate the density. Here re-clustering is needed to find the frequent item sets to retrieve the data easily. The support and confidence gives the efficiency and improve the performance of data exchange. Where it shows that the partitioned data is reduced to some levels based on the frequent data for the dataset and we represent the graph which helps for the easily understanding of the client about the information needed of the given dataset. The results are shown below in dataset values, cluster values and in graph values are represented First, the focus of data exchange is on scalability, where instead of generating an arbitrary number of super-entities that may exponentially increase according the number of source tuples, introduce and employ the relation and tuple tree structure that can be used to formally define the similarity between source tuples and target relations.

This approach not only eliminates the need for pruning a large collection of super-entities, but also allows reusing scripts that are

already generated for tuple trees with the same structure. Second, novel techniques are proposed to compare tuple trees of source and relation trees of target. Third, the script reusing technique propose ensures scalability of processing. It report an extensive set of experiments to show how data exchange scenarios outperforms existing works. Finally security is provided to the target schema through l-diversity method. The above illustrates the schema mapping exchange the data from source to target without occur any duplication of data as well as data exchange. Implement script reusing technique in the phase of generating transformation scripts, it is possible to reuse the scripts generated for tuples in each scripts. The utility based cache preventing data optimize the original data and providing security through l-diversity.

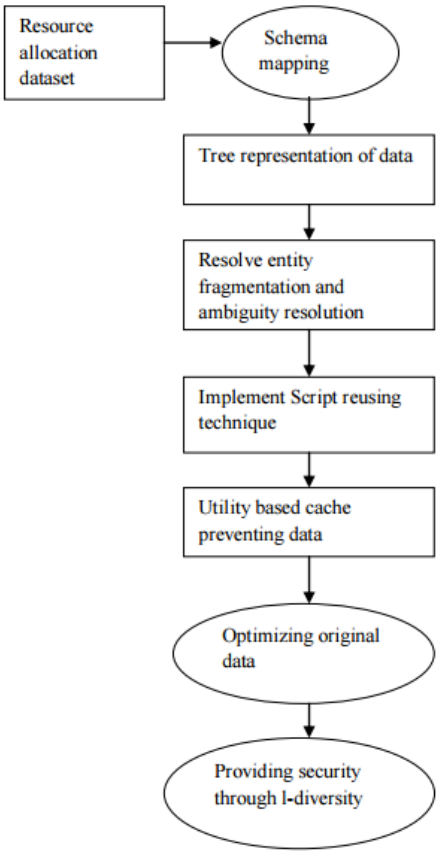
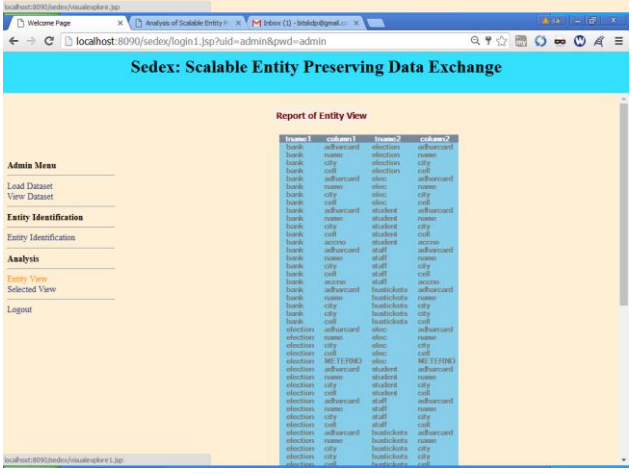
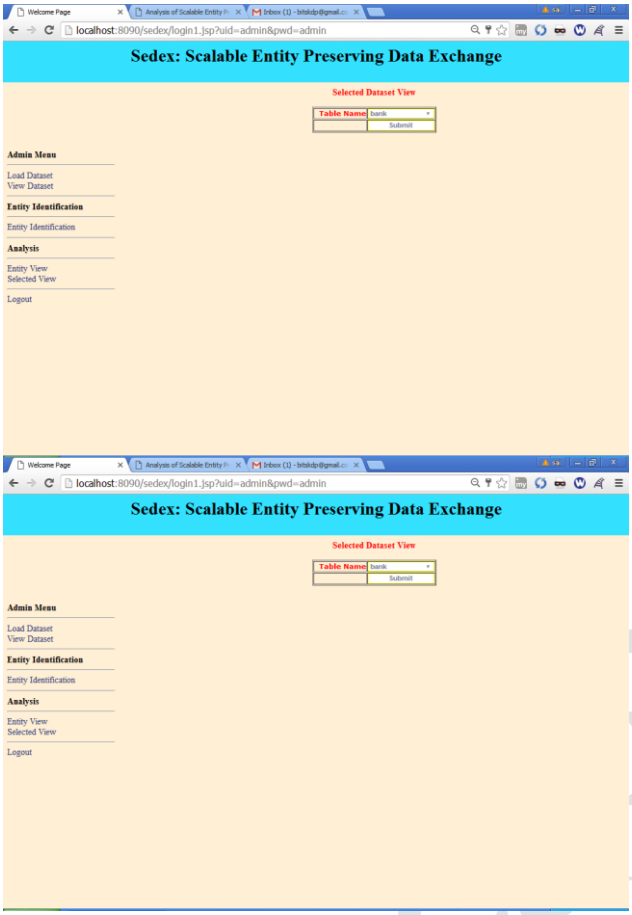


Fig 1 System Architecture

In this paper, a mapping techniques of class based approach and transformation scripts of data exchange from the implementations of a generalization relation and entity preserving data exchange bridges the gap between data level and schema level approaches and it avoids entity fragmentation and resolves ambiguous data exchange scenarios.





IV. CONCLUSION

We have investigated the use of clustering to group all the frequent information to assist and prioritize their analysis.. Here we are discussing as preserving approach entity pattern mapping for data exchange in which the focus is on preserving source entities in the target no matter which class they belong to in the source. As we showed for pattern mataching can directly generate the expected solution as a desirable solution for data exchange. In enhancement, as data is stored in Hadoop or bigdata. Bigdata with large number of data is stored and finding map & Reduce. We created abstract representations of source and target by forming a tree structure of source entities and target relations. The decision trees used in proposed scheme gives the information based on the user choice by considering the frequent itemset, that the particular information to be retrieve based on after checking conditions on each and every attribute and finally merging the data into one thereby we will get the desired information. Then, using tree similarity techniques which work based on finding distance functions between trees, the best relation trees matching the source entities were identified. Different algorithms for improving data efficiency and finding job schedule and word count in Hadoop are to be used for the purpose of better performance and consistency in the matters of retrieving the any type of information from the databases.

REFERENCE

[1] Wu, X., Zhu, X., Wu, G. Q., and Ding, W.: Data mining with big data. Knowledge and Data Engineering, IEEE Transactions on, 26(1), 97-107 (2014)
[2] Popa, L., Velegrakis, Y., Hernandez, M. A., Miller, R. J., and Fagin, R.: Translating Web data. Proc. Very Large Data Bases, 598-609(2002)
[3] Miller, R. J., Haas, L. M., and Hernadez, M. A.: Schema mappings query discovery. Proc. Very Large Data Bases, 77-88 (2000)
[4] Fagin, R., Kolaitis, P. G., Miller, R. J., and Popa, L.: Data exchange: semantics and query answering. Theoretical Computer Science, 336(1), 89-124 (2005)
[5] Fagin, R., Kolaitis, P. G., and Popa, L.: Data exchange: getting to the core. ACM Transactions on Database Systems, 30(1), 174-210 (2005)
[6] Gottlob, G., and Nash, A.: Efficient core computation in data exchange. Journal of the ACM, 55(2), 9. (2008)
[7] Pichler, R., Savenkov, V.: Towards practical feasibility of core computation in data exchange. Theoretical Computer Science,411(7-9), 935-957 (2010)
[8] Marnette, B., Mecca, G., Papotti, P.: Scalable data exchange with functional dependencies. Proc. Very Large Data Bases, 3(1-2),105-116 (2010)
[9] Qian, L., Cafarella, M. J., and Jagadish, H. V.: Sample-drivenschema mapping. Proc. ACM SIGMOD International Conferenceon Management of Data, 73-84 (2012)
[10] Alexe, B., ten Cate, B., Kolaitis, P. G., Tan, W.: EIRENE:Interactive design and refinement of schema mappings via dataexamples. Proc. VLDB Endowment, 4(12), 1414-1417 (2011)
11] Alexe, B., Hernndez, M., Popa, L., Tan, W. C.: MapMerge:correlating independent schema mappings. The VLDB Journal,21(2), 191-211 (2012)
[12] Arocena, P. C., Glavic, B., Ciucanu, R., and Miller, R. J.: TheiBench integration metadata generator. Proc. Very Large DataBases, 9(3), 108-119 (2015)
[13] Barbosa, D., Mendelzon, A. O., Keenleyside, J., and Lyons, K.A.: ToXgene: An extensible template-based data generator forXML. Proc. WebDB, 49-54 (2002)
[14] Arenas, M., Prez, J., Reutter, J., and Riveros, C.: Compositionand inversion of schema mappings. ACM SIGMOD Record,38(3), 17-28 (2010)
[15] Fagin, R., Kolaitis, P. G., Popa, L., and Tan, W. C.: Reverse dataexchange: coping with nulls. ACM Transactions on DatabaseSystems, 36(2), 11 (2011)
[16] Bunge, M.v.: Treatise on Basic Philosophy: the Furniture of theWorld. Boston, MA: Reidel (1977)
[17] Mecca, G., Papotti, P., Raunich, S., and Santoro, D.: What isthe IQ of your Data Transformation System?. Proc. Informationand knowledge management, 872-881 (2012)
[18] Sekhavat, Y. A., and Parsons, J.: SESM: semantic enrichment ofschema mappings. Proc. ICDE International Workshop on DataEngineering Meets Semantic Web, 7-12 (2013)
[19] Sekhavat, Y. A., and Parsons, J.: EDEX: Entity Preserving DataExchange. Proc. DATA, 221-229, (2013)
[20] Parsons, J., W and, Y.: Emancipating instances from thetyranny of classes in information modeling. ACM Transactionson Database Systems, 25(2), 228-268 (2000)