Core Solution Computing Algorithm of Web Data Exchange

Yuhang Ji[1,2(✉)], Gui Li[1,2], Zhengyu Li[1,2], Ziyang Han[1,2], and Keyan Cao[1,2]

[1] Computer Engineering and Applications, Shenyang Jianzhu University, Shenyang 110168, China
syjzjyh@163.com

[2] Faculty of Information and Control Engineering, Shenyang Jianzhu University, Shenyang 110168, China

Abstract. Traditional Web data exchange research usually focuses on designing transformation rules but ignores the processing of the actual generated target data instances. Since the data instance is highly correlated with the schema and there are many duplicate elements in the source data instance, there is redundancy in the actual generated target data instance. In order to generate a target data instance solution that does not contain redundancy, under a given source-to-target exchange rule, a unified integration schema is designed firstly, and then, the instance block mechanism is introduced to analyze three mapping relationships of single homomorphism, full homomorphism and the isomorphism among the initial generated target data instances. According to the mapping relationship, three methods of instance selection, which are more compact, more informative and equivalence class processing, are proposed to remove redundant data instance in target data set and generate the core solution of target data instance. The experiment uses the data from the China Land Market Network to evaluate the performance of the data exchange core solution algorithm.

Keywords: Large web data · Data exchange · Redundant data processing · Core solution · Homomorphism relation · Schema integration

1 Introduction

Data exchange is very important in multi-source data integration. The original data exchange problem [1] was proposed by Fagin et al., the paper illustrates that data conversion typically takes source data as input and select the source data by a set of mapping rules (also known as tuple generation dependencies) to transform it into a target data set that satisfies a given schema mapping rule. On this basis, Web data exchange can be carried out at two levels, schema layer and instance layer [2]. The main work of schema layer is to design an accurate and complete set of mapping rules according to the attribute correspondence between source and target. The instance layer obtains and generates the target data set from multiple Web data sources according to the set of schema mapping rules between the given source and target database, and

W. Ni et al. (Eds.): WISA 2019, LNCS 11817, pp. 113–118, 2019.
https://doi.org/10.1007/978-3-030-30952-7_13

performs repeated data processing on the target data set to remove redundant data instances. However, traditional research on Web data exchange usually focuses on designing schema exchange rules, while ignoring the processing of the actual generated target data instance. In the data exchange scenario shown in Fig. 1, table A, B, C, and D represent the source database tables from different web sources. Table T1 and T2 represent the target database tables. The attribute correspondence of the source-to-target database tables and the given mapping rules are shown in Fig. 2.

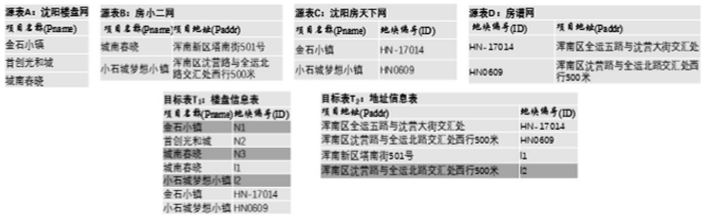

Fig. 1. Source and target profile data

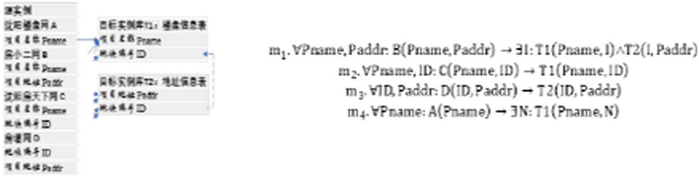

Fig. 2. Attribute correspondence and mapping rules

The target data set of the target table T1, T2 in Fig. 1 can be obtained by given schema mapping rules, and these target data sets are called the canonical universal solution [2]. They basically satisfy the given mapping rules, but contain Multiple redundant tuples (gray background part). In this paper, T1 is obtained on the basis of the initial target data instance sets and T2 does not contain redundant target data instance sets, that is, the part with the white background in the target table T1 and T2. This part can be called the core universal solution. [2].

The contribution can be sum up as follows:

- We design an accurate integration schema and given the initial mapping rules.
- Three instance selection methods based on homomorphic relation are proposed to calculate data exchange core solution. Finally, a detailed experiment and theoretical analysis of the data exchange algorithm are carried out with the online data set, and the experimental results verify the performance of the algorithm in this paper.

2 Related Work

With the development of Web technology and the popularization of its application, Web data has become the main data source in various fields. Pichler et al. [3] started from the perspective of schema mapping, through the optimization of the mapping rules and converted them into executable scripts to calculate the kernel solution. However, when faced with large data exchange scenarios, this technique may result in a large amount of redundant data in the target database. In literature [6], it is considered that data exchange based on schema-level information only limits the ability to express semantics in data exchange, and cannot solve some fuzzy data exchange scenarios. In literature [9] and [10], data instances are used to reconstruct schema mapping, and constraints in the pattern are used to find natural associations and it hopes to find the kernel solution through better examples and mapping rules. The main difficulty with this method is the selection of data instances, which may be redundant due to the fact that different data instances may describe the same thing.

3 Methodology

3.1 The Achievement of the WEB Multisource Integration Mode

Because the relevance between data and mode is very high in the process of exchanging the web data, only when the mode integrates [2] and designs an initial exchange rule, the living example can be exchanged from source data into target data. The process of integration mode can be divided into two phases:

The Definition Phase of Integration Mode. Firstly, the data source to be exchanged and integrated should be determined. Be directed against different data sources, we should analyze their mode information. Analyze the universality and differences between the output modes of various part databases. Define the formal description of integration mode on the basis of meeting the needs of user data and obtain the all information which can be used in the follow-up procedures of integration mode.

The Definition Phase of Matching Relation Table. On the basis of the formal description of integration mode, we define the matching relation tables which include the matching relationship between output mode of source databases and integration mode of target databases in data table names, attribute names and operation names. Find certain mapping relationships which local in two different modes' elements. Input two modes as parameters and the output result is the mapping relationship between then which is the initial mapping rule set.

3.2 The Selection Method of Data Instance

Firstly, find all the instance block sets which content transfer rules of data. Obtain the irredundant instance block sets by using the homomorphism relationship between instance blocks to delete redundant instance blocks. Then, choose the instance blocks with higher accuracy to produce that by calculating the accuracy of data in instance blocks. According to the instance blocks, the basic features of core and popular solution can be defined. Choose and output the final target instance data.

Because there are different homomorphism relationships in instance blocks, this paper will use homomorphism to define what the "redundant" instance block is.

The Classification in Homomorphism. Given two instance set J and J', mapping $h: J \to J'$.When $J \to h(J), J' \to h(J')$, if $J * J' \to h(J) * h(J')$, we call that the mapping h is the homomorphism from J to J'. Above all, there are two main reasons due to redundant instance blocks. The first one is that there is the epimorphism relationship between w and w' which makes w' more compact than w, and w is the redundant block which expressed by $w \prec w'$. The second one is that even we exclude the tuples included many uncertainties, the instance block w may still produce other instance block w' which exists with single homomorphism relationship by using other exchange rules and assignment, we call that w' has more information than w, w is the redundant block which expressed by $w \prec w'$.

Through the above steps, the most redundancies in initial solutions can be removed, but not enough to produce core solutions. When two biggest instance blocks are isomorphism to each other, the core solution only need to consider one of them.

In order to produce the instance block sets which can accurately calculate core solutions, the accuracy of each instance block w must be calculated first. Remove the instance blocks with lower accuracy until the change of accuracy $R(W) = \sum_{v \in v(w)} \frac{p(v)}{n}$ in each instance blocks are smaller than the given standard value $\left(P(v) = \frac{\sum_{i=1}^{v_j} Simv_i, v_i}{v_i}\right)$. Then, choose the instance blocks with higher accuracy to make sure that there is no isomorphism relationship between instance blocks of the set. Finally, take advantage of these instance block sets to produce the core solutions. The accuracy of instance blocks can be calculated by the following formula.

4 Experiments

4.1 Experimental Setting and Dataset

The experimental data set used in this paper is from China Land and Market Network (www.landchina.com). To evaluate the performance of the algorithm, the real-world data is difficult to present all the problems, so the artificial data set is constructed by using the above data set. Three of these classified attribute data are chose to do the experiment. The information description of the data set includes 18052 estate records, 25308 item indicia records and 62371 address records.

In order to compare the superiority of this method and other similar works, we design the source databases which include 100k, 250k, 500k and 1M tuples. Compare our method with the computational algorithm of core solution in literature [3] and literature [4]. t1, t2 and t3 are used separately to represent our method and other two computational algorithms. Test the running time of our method and others aimed at core solution computing problems on the large instance tuples. We design five source databases with 10k data which are all from the real data set to prove through experiments that the target instances produced by core solution is less and better in quality than by standard solution. Every source database includes different degrees of

"redundant" and the range is 0%–40%. We do the specific experiments with eight different mapping scenes. S1–S4 does not include self-join. SJ1–SJ4 includes self-join.

4.2 Experimental Results

As shown in Fig. 3, it is obvious that the time of the target core solutions in computing large data set is less and the efficiency of that is higher by using our method.

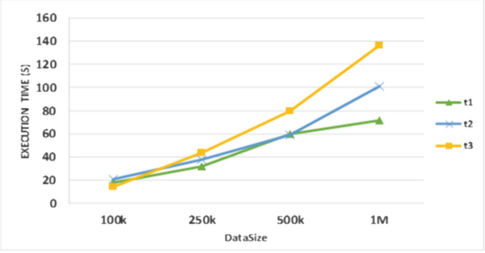

Fig. 3. Performance testing

By comparing the tuple numbers produced in the target database, the reduced generating tuple numbers percentage of the core solution compared with the universal solution is obtained. The Fig. 4 left shows the four project results of the target which do not include self-join and the core solution is more compact than the universal solution in all cases. As redundancy increases, this case becomes more apparent. Two hypothetical scenarios S1 and S2 follow the trend but not as significant as the two coverage scenarios S3 and S4, because the design of tgds in S1 and S2 often generates many duplicate tuples in the solutions and these tuples are deleted by core scripts and universal scripts. Figure 4 right shows the reduced percentage of four self-join solutions. Except SJ_1, the core solution is more compact than the universal solution in all cases. No *null* value is generated in the solution which the universal solution and the core solution coincide, because SJ_1 is the complete mapping.

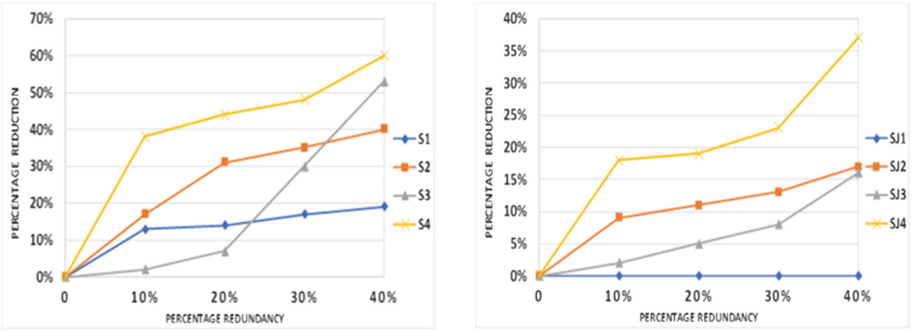

Fig. 4. Reduction of redundancy

5 Conclusion

In this paper, we proposed a data conversion kernel calculation method based on homomorphic relationship between target database instances. We used data from the China Land and Market Network (www.landchina.com) to build datasets and to evaluate the performance of the algorithm, three of these classified attribute data are chose to do the experiment. Experimental results prove the effectiveness of our method.

References

1. Fagin, R., Kolaitis, P.G., Miller, R.J., Popa, L.: Data exchange: semantics and query answering. In: Calvanese, D., Lenzerini, M., Motwani, R. (eds.) ICDT 2003. LNCS, vol. 2572, pp. 207–224. Springer, Heidelberg (2003). https://doi.org/10.1007/3-540-36285-1_14
2. Fagin, R., Kolaitis, P.G., Popa, L.: Data exchange: getting to the core. ACM Trans. Database Syst. **30**(1), 174–210 (2005)
3. Pichler, R., Savenkov, V.: Towards practical feasibility of core computation in data exchange. Theoret. Comput. Sci. **411**(7), 935–957 (2010)
4. Gottlob, G., Nash, A.: Data exchange: computing cores in polynomial time. In: ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. ACM (2006)
5. Cabibbo, L.: On keys, foreign keys and nullable attributes in relational mapping systems. In: International Conference on EDBT. DBLP (2009)
6. Sekhavat, Y.A., Parsons, J.: SEDEX: scalable entity preserving data exchange. IEEE Trans. Knowl. Data Eng. **28**(7), 1878–1890 (2016)
7. Mecca, G., Papotti, P., Raunich, S.: Core schema mappings: scalable core computations in data exchange. Inf. Syst. **37**(7), 677–711 (2012)
8. Cai, D., Hou, D., Qi, Y., Yan, J., Lu, Y.: A distributed rule engine for streaming big data. In: Meng, X., Li, R., Wang, K., Niu, B., Wang, X., Zhao, G. (eds.) WISA 2018. LNCS, vol. 11242, pp. 123–130. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-02934-0_12
9. Arocena, P.C., Glavic, B., Ciucanu, R., Miller, R.J.: The iBench integration metadata generator. Proc. Very Large Data Bases **9**(3), 108–119 (2015)
10. Dong, X.L., Berti-Equille, L., Srivastava, D.: Integrating conflicting data: the role of source dependence. Proc. VLDB Endowment **2**(1), 550–561 (2018)
11. Han, Z., Jiang, X., Li, M., Zhang, M., Duan, D.: An integrated semantic-syntactic SBLSTM model for aspect specific opinion extraction. In: Meng, X., Li, R., Wang, K., Niu, B., Wang, X., Zhao, G. (eds.) WISA 2018. LNCS, vol. 11242, pp. 191–199. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-02934-0_18