# The Interaction between Schema Matching and Record Matching in Data Integration (Extended Abstract)

Binbin Gu
Soochow University,
Suzhou, China
Email: gu.binbin@hotmail.com

Zhixu Li
Soochow University,
Suzhou, China
Email: zhixuli@suda.edu.cn

Xiangliang Zhang
KAUST,
Jeddah, Saudi Arabia
Email: xiangliang.zhang@kaust.edu.sa

An Liu
Soochow University,
Suzhou, China
Email: anliu@suda.edu.cn

Guanfeng Liu
Soochow University,
Suzhou, China
Email: gfliu@suda.edu.cn

Kai Zheng
Soochow University,
Suzhou, China
Email: kaizheng@suda.edu.cn

Lei Zhao
Soochow University,
Suzhou, China
Email: zhaol@suda.edu.cn

Xiaofang Zhou
The University of Queensland,
Brisbane QLD 4072, Australia
Email: zxf@itee.uq.edu.au

## I. Background and Motivation

Schema Matching (SM) [3] and Record Matching (RM) [1] are two necessary steps in integrating multiple relational tables of different schemas, where SM unifies the schemas and RM detects records referring to the same real-world entity. The two processes have been thoroughly studied separately, but few attention has been paid to the interaction of SM and RM. In this work we find that, even alternating them in a simple manner, SM and RM can benefit from each other to reach a better integration performance (i.e., in terms of precision and recall). Therefore, combining SM and RM is a promising solution for improving data integration [2].

For instance, assume a start-up linked key attribute-pairs $(\mathtt{Product}, \mathtt{Product})$ between the two tables in Fig. 1 (a), at the first RM step, we may identify $(t_1 \leftrightsquigarrow s_1)$ and $(t_2 \leftrightsquigarrow s_2)$ as linked records as they share the same Product values. We then identify $(\mathtt{Weight}, \mathtt{WT})$ as linked attribute-pair given that the two linked records share the same value under the two attributes. We repeat this process iteratively until no more attributes or records can be linked. Finally, we will have all the four attribute-pairs and six record-pairs be correctly linked as demonstrated in Fig. 2. By contrast, traditional methods perform SM and RM in only one run, which as a result introduce $(\mathtt{Ex-Memory} \leftrightsquigarrow \mathtt{ROM})$ and $(t_4 \leftrightsquigarrow s_8)$ as wrong matches, and also miss pairs $(\mathtt{Size} \leftrightsquigarrow \mathtt{Screen\ Size})(\mathtt{t_3} \leftrightsquigarrow \mathtt{s_3}), (\mathtt{t_4} \leftrightsquigarrow \mathtt{s_4}), (\mathtt{t_5} \leftrightsquigarrow \mathtt{s_5})$ and $(t_6 \leftrightsquigarrow s_6)$ as matched pairs with similarities and thresholds given in Fig. 1(b)(c). We now describe a basic interaction workflow between SM and RM with an example scenario in Fig. 2.

*Example 1:* Initially, we have $\mathcal{P}_0^S = \{(\mathtt{Product} \leftrightsquigarrow \mathtt{Product})\}$, according to which we can match $(t_1 \leftrightsquigarrow s_1)$ and $(t_2 \leftrightsquigarrow s_2)$. Then, we find that the two matched records share the same values under $(\mathtt{WT}, \mathtt{Weight})$ and $(\mathtt{SIZE}, \mathtt{Screen})$. Thus, $(\mathtt{WT} \leftrightsquigarrow \mathtt{Weight})$ and $(\mathtt{SIZE} \leftrightsquigarrow \mathtt{Screen})$ can be our newly-linked attribute-pairs. Until now, we would have three attribute-pairs, according to which we can find a new record-pair $(t_4 \leftrightsquigarrow s_3)$, given that the three matching-attribute-pairs support this record-pair. Next, since $t_4[\mathtt{CAMERA}]$ equals to $s_3[\mathtt{BackCam}]$ rather than $s_3[\mathtt{FontCam}]$, we may have $(\mathtt{CAMERA} \leftrightsquigarrow \mathtt{BackCam})$. We continue with RM and SM alternatively in this way to have $(t_5 \leftrightsquigarrow s_4)$ and $(\mathtt{ROM} \leftrightsquigarrow \mathtt{Memory})$.

Given the above intuition, we study the interaction between SM and RM, by performing them alternately for data integration.

## II. Problem Statement

There are several crucial issues in the interaction workflow. Firstly, the way of estimating the matching likelihood of an attribute-pair (or a record-pair) is the key factor to ensure the matching quality. The matching likelihood between two records depends on two aspects, i.e., the number of linked attribute-pairs that support the matching, and the ability of the linked attribute-pair in recognizing matching-record-pairs.

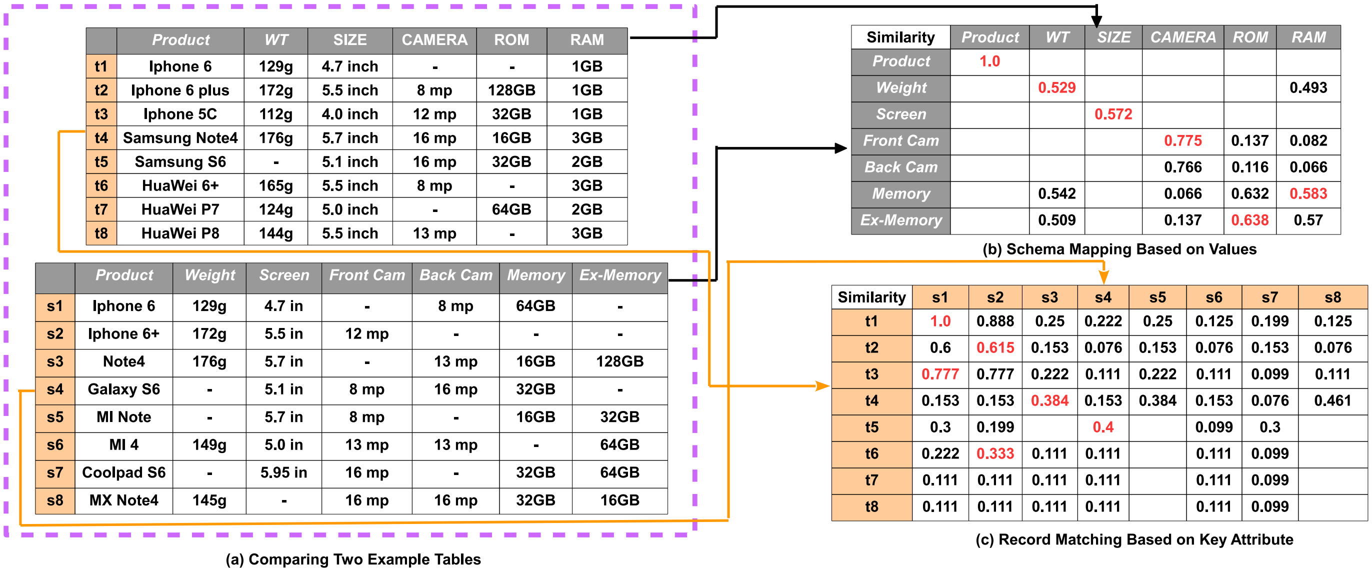
Second, "semantic drift" problem should be controlled for preventing the mistake magnification from an SM (or RM) step to the following RM (or SM). The linking decisions made at each SM (or RM) step based on temporary RM (or SM) results should be validated.

Last but not the least, the large overhead produced by comparing a large number of value pairs should be reduced.

## III. Summary of Our Approaches

To estimate the matching likelihood between two records, we firstly define the ability of recognizing matching-record-pair of an attribute called "$IdC$ socre". Secondly, we combine the contributions from multiple-pairs to the calculation of matching likelihood of two related records. However, the attributes are not always independent which makes the calculation intractable. To solve this problem, we first assume that all the IdC of attributes are independent such that a linear model can be used to calculate the matching likelihood, and then we compensate for the dependence between attributes in the model by introducing a damping factor.

To control the semantic drift problem, we validate the newly-linked records and newly-linked attributes separately to prevent semantic drift from happening. After each RM step, we identify risky record-pairs by checking the unbiased



|  | Product | WT | SIZE | CAMERA | ROM | RAM |
|---|---|---|---|---|---|---|
| t1 | Iphone 6 | 129g | 4.7 inch | - | - | 1GB |
| t2 | Iphone 6 plus | 172g | 5.5 inch | 8 mp | 128GB | 1GB |
| t3 | Iphone 5C | 112g | 4.0 inch | 12 mp | 32GB | 1GB |
| t4 | Samsung Note4 | 176g | 5.7 inch | 16 mp | 16GB | 3GB |
| t5 | Samsung S6 | - | 5.1 inch | 16 mp | 32GB | 2GB |
| t6 | HuaWei 6+ | 165g | 5.5 inch | 8 mp | - | 3GB |
| t7 | HuaWei P7 | 124g | 5.0 inch | - | 64GB | 2GB |
| t8 | HuaWei P8 | 144g | 5.5 inch | 13 mp | - | 3GB |

|  | Product | Weight | Screen | Front Cam | Back Cam | Memory | Ex-Memory |
|---|---|---|---|---|---|---|---|
| s1 | Iphone 6 | 129g | 4.7 in | - | 8 mp | 64GB | - |
| s2 | Iphone 6+ | 172g | 5.5 in | 12 mp | - | - | - |
| s3 | Note4 | 176g | 5.7 in | - | 13 mp | 16GB | 128GB |
| s4 | Galaxy S6 | - | 5.1 in | 8 mp | 16 mp | 32GB | - |
| s5 | MI Note | - | 5.7 in | 8 mp | - | 16GB | 32GB |
| s6 | MI 4 | 149g | 5.0 in | 13 mp | 13 mp | - | 64GB |
| s7 | Coolpad S6 | - | 5.95 in | 16 mp | - | 32GB | 64GB |
| s8 | MX Note4 | 145g | - | 16 mp | 16 mp | 32GB | 16GB |

(a) Comparing Two Example Tables

| Similarity | Product | WT | SIZE | CAMERA | ROM | RAM |
|---|---|---|---|---|---|---|
| Product | 1.0 |  |  |  |  |  |
| Weight |  | 0.529 |  |  |  | 0.493 |
| Screen |  |  | 0.572 |  |  |  |
| Front Cam |  |  |  | 0.775 | 0.137 | 0.082 |
| Back Cam |  |  |  | 0.766 | 0.116 | 0.066 |
| Memory |  | 0.542 |  | 0.066 | 0.632 | 0.583 |
| Ex-Memory |  | 0.509 |  | 0.137 | 0.638 | 0.57 |

(b) Schema Mapping Based on Values

| Similarity | s1 | s2 | s3 | s4 | s5 | s6 | s7 | s8 |
|---|---|---|---|---|---|---|---|---|
| t1 | 1.0 | 0.888 | 0.25 | 0.222 | 0.25 | 0.125 | 0.199 | 0.125 |
| t2 | 0.6 | 0.615 | 0.153 | 0.076 | 0.153 | 0.076 | 0.153 | 0.076 |
| t3 | 0.777 | 0.777 | 0.222 | 0.111 | 0.222 | 0.111 | 0.099 | 0.111 |
| t4 | 0.153 | 0.153 | 0.384 | 0.153 | 0.384 | 0.153 | 0.076 | 0.461 |
| t5 | 0.3 | 0.199 |  | 0.4 |  | 0.099 | 0.3 |  |
| t6 | 0.222 | 0.333 | 0.111 | 0.111 |  | 0.111 | 0.099 |  |
| t7 | 0.111 | 0.111 | 0.111 | 0.111 |  | 0.111 | 0.099 |  |
| t8 | 0.111 | 0.111 | 0.111 | 0.111 |  | 0.111 | 0.099 |  |

(c) Record Matching Based on Key Attribute

Fig. 1. Two Example Tables for Integration (a) and the Integration Results with Previous Methods ((b) and (c))

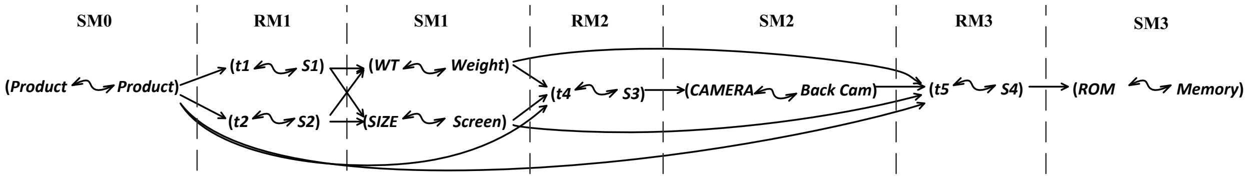


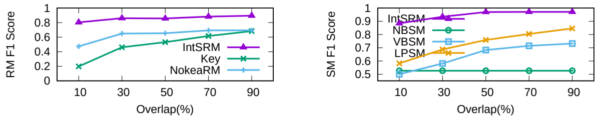
Fig. 2. Example Interaction Workflow of SM and RM for Integrating Tables in Fig. 1

variance of the similarity between their value pairs under various attribute-pairs, while after each SM step, we identify outlier attribute-pairs by applying cross-validation techniques to validate all the linked attributes.

To optimize the time cost of comparing multiple attribute value pairs, we extend the q-gram index [4] to multiple pairs of attributes scenario, and split potential matched record-pairs between the two tables into (possibly overlapped) blocks so that matching-record-pairs are only identified within every block. Then we further block record-pairs by computing the upper and lower-bound of the matching likelihood of two record.

## IV. Evaluation

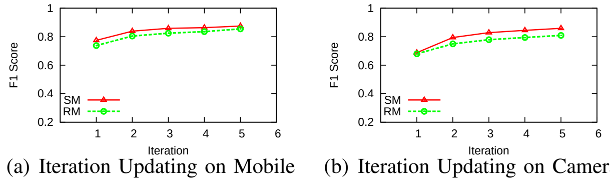
**(1) $F_1$ Comparison for RM and RM:** Since the overlap ratio of the records between two tables has a great influence on the integration quality, we conduct our comparison experiments at various overlap ratios (from 10% to 90%) on the three data sets. As demonstrated in Fig. 3, our method IntSRM always reaches the highest $F-1$ score for both RM and SM.



(a) RM on Mobile (b) SM on Mobile

Fig. 3. $F_1$-score of RM and SM Methods respectively on the Camara Dataset

**(2) Iterative Updating:** As demonstrated in Fig. 4(a)(b), the quality of RM and SM can also make a further improvement and they can hold steady with a satisfied result as the iteration goes.



(a) Iteration Updating on Mobile (b) Iteration Updating on Camera

Fig. 4. Quality Improvement with Interaction

## Acknowledgements

This research is partially supported by Natural Science Foundation of China (Grant No. 61402313, 61303019, 61472263, 61572336), the King Abdullah University of Science and Technology, the AustralianResearch Council (Grants No. DP120102829), the Postdoctoral scientific research funding of Jiangsu Province (No. 1501090B) and the National Postdoctoral Funding (No. 2015M581859, 2016T90493).

## References

[1] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate record detection: A survey. *TKDE*, 19(1):1–16, 2007.

[2] B. Gu, Z. Li, X. Zhang, A. Liu, G. Liu, K. Zheng, L. Zhao, and X. Zhou. The interaction between schema matching and record matching in data integration. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):186–199, 2017.

[3] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *the VLDB Journal*, 10(4):334–350, 2001.

[4] C. Xiao, W. Wang, and X. Lin. Ed-join: an efficient algorithm for similarity joins with edit distance constraints. *PVLDB*, 1(1):933–944, 2008.