Attribute Value Matching by Maximizing Benefit

Fengfeng Fan[1,2](✉) and Zhanhuai Li[1,2]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
fanfengfeng@mail.nwpu.edu.cn

[2] Key Laboratory of Big Data Storage and Management, Ministry of Industry and Information Technology, Northwestern Polytechnical University, Xi'an, China

Abstract. Attribute value matching (AVM) identifies equivalent values that refer to the same entities. Traditional approaches ignore the weights of values in itself. In this demonstration, we present AVM-LB, Attribute Value Matching with Limited Budget, that preferentially matches the *hot* equivalent values such that the maximal benefit to data consistency can be achieved by limited budget. By defining a `rank` function and greedily matching the hot equivalent values, the AVM-LB enables users to interactively explore the achieved benefit to data consistency.

Keywords: Attribute Value Matching · Entity resolution · Hot data
Data cleaning · Big Data

1 Introduction

Due to typographical errors, aliases and abbreviations [1,4], the same real-world entities may take several distinct representations across data sources, and such inconsistencies may severely distort the results of data analysis. Hence it is necessary to match and merge those equivalent values by a process called Attribute Value Matching or AVM [3]. Due to the large data size and limited budget, it is a very challenging task to identify *all* of underlying equivalents, thus it is preferred to employ a pay-as-you-go approach [5] to identify the equivalent attribute values. However, existing approaches ignore the fact that inconsistencies between frequently accessed *hot* attribute values will bring more distortion to data analysis and matching the *hot* equivalent values will bring more benefit to data consistencies. In this paper, we propose a demo, denoted by AVM-LB, which takes the *matching probability* and *data hotness* into consideration, and interactively explores the achieved benefit by limited budget. To our knowledge, AVM-LB is the first demonstration that incorporates the *data hotness* into data cleansing practice. Our contributions can be summarized as follows:

1. AVM-LB provides a `rank` function, which ranks the candidates of value pairs for resolving, based on the *matching probability* and *hotness*.

Supported by organization x.

C. Woo et al. (Eds.): ER 2018 Workshops, LNCS 11158, pp. 22–25, 2018.
https://doi.org/10.1007/978-3-030-01391-2_5

2. Based on the *matching relationship* and the *data hotness*, a benefit metric is devised to quantify the improvement to data consistency.
3. AVM-LB enables users to interactively explore the achieved benefit to data consistencies with limited budget.

2 System Overview

AVM-LB is composed of two components: Benefit Maximization and Performance Evaluation.

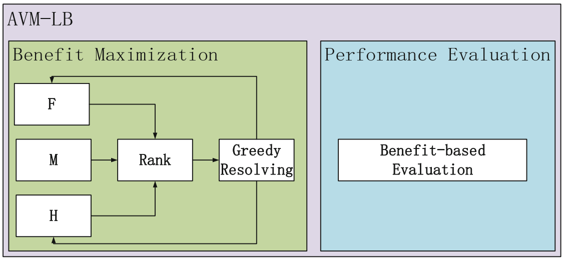


Fig. 1. System overview

2.1 Benefit Maximization

As the Fig. 1 shows, the rank function takes three matrices as input: filter matrix **F**, matching probability matrix **P** and hotness matrix **H**.

Filter Matrix: In AVM-LB, filter matrix **F** maintains the current states between attribute values. A snapshot of **F** is shown in Eq. 1: "1" for matches, "−1" for non-matches, "0" for unknowns, and "*" for links that can be deduced by symmetry.

$$\mathbf{F} = \begin{bmatrix} * & 1 & 0 & 0 \\ * & * & -1 & 0 \\ * & * & * & 1 \\ * & * & * & * \end{bmatrix} \tag{1}$$

As the matching process goes on, more 0-labeled cells will be replaced by either "1" or "−1", depending on the matching results, until no more 0-labels is available or all the budget run out.

Matching Probability: Matching probability matrix **M**, maintains the matching probabilities between attribute values, with $\mathbf{M}[i,j] \leftarrow \mathsf{P}(y_i \cong y_j)$. For simplicity, we approximate the matching probability $\mathsf{P}(y_i \cong y_j)$ by a similarity function $\mathsf{sim}(y_i, y_j)$, which can either be a simple string similarity measurement or some sophisticated metric, e.g., [2].

Hotness: Hotness often reveals the attribute value's importance in data analysis, and it may be a function of the timeliness, occurrences, or access frequencies. For similarity, we estimate the *hotness* of attribute values by their frequencies.

We define the hotness for any attribute value pair $\langle y_i, y_j \rangle$ by Eq. 2:

$$\mathsf{hot}(y_i, y_j) = \mathsf{freq}([y_i]) \cdot \mathsf{freq}([y_j]) \tag{2}$$

where the equivalent class $[y_i]$ denotes the set of attribute values co-referring to the same entity with y_i, and $\mathsf{freq}(\cdot)$ records the frequencies of attribute values. Hotness matrix $\mathbf{H}$, maintains the hotnesses for attribute value pairs, i.e., $\mathbf{H}[i, j] \leftarrow \mathsf{hot}(y_i, y_j)$.

Rank: AVM-LB ranks the value pairs by a integrated scores, which is defined by Eq. 3:

$$\mathsf{rank}(y_i, y_j) = \bar{\mathbf{F}}[i, j] \cdot \mathbf{M}[i, j] \cdot \mathsf{sigmoid}(a \cdot \mathbf{H}[i, j] + b) \tag{3}$$

where $\bar{\mathbf{F}}$, the negation of $\mathbf{F}$, is used to filter out the resolved value pairs, the transformation from hotness into weight is provided by $\mathsf{sigmoid}$ function, in which $a \geq 0$ and b as two tuning parameters, and the matrix **Rank** maintains the integrated scores, i.e., $\mathbf{Rank}[i, j] \leftarrow \mathsf{rank}(y_i, y_j)$.

Finally with limited budget Ks, AVM-LB greedily matches the equivalents based on the value of $\mathbf{Rank}[i, j]$.

2.2 Performance Evaluation

AVM-LB evaluates the performance by $\mathsf{benefit}$, which is defined by Eq. 4:

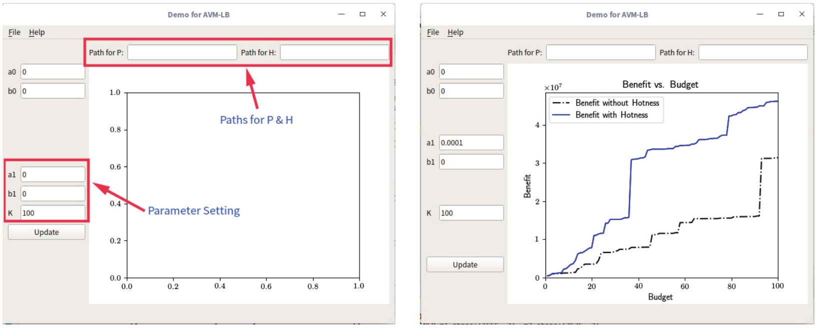
$$\mathsf{benefit}(y_i, y_j) = \mathsf{I}(y_i, y_j) \cdot \mathsf{hot}(y_i, y_j) \tag{4}$$

where the indicator function $\mathsf{I}(\cdot)$ will return 1 for $y_i \cong y_j$, and 0 for otherwise. Intuitively, high $\mathsf{benefit}$ will bring big improvement to the probability of receiving consistent view of data for random queries.

For demonstrative purpose, we match the `Journal` values across two public available datasets, DBLP[1] and CiteSeer[2], in which $1,636,497$ and $45,783$ records are analyzed, and $1,666$ and $3,833$ distinct `Journal` values are extracted respectively. We construct the matching probability matrix $\mathbf{P}$ and hotness matrix $\mathbf{H}$ following the method in [2] and the definition of Eq. 2 respectively.

Figure 2(a) shows the startup user-interface, in which paths for dump file of $\mathbf{P}$ and $\mathbf{H}$ needs to be specified. After setting the valid paths for $\mathbf{P}$ and $\mathbf{H}$, we can interactively explore the achieved benefits by tuning parameters of rank function. For example, Fig. 2(b) shows the accumulated $\mathsf{benefit}$ with different budget by different rank function, in which the dashed curve ignores the hotness by setting parameter $a_0 = 0$, while the solid curve fine-tunes the weight of attribute value pairs by setting parameter $a_1 = 0.0001$ and $b_1 = 0$. It can be observed that by tuning parameters, AVM-LB allow us to interactively explore and visualize the $\mathsf{benefit}$ to data consistency with limited budget.

[1] http://dblp.uni-trier.de/.
[2] https://www.cs.purdue.edu/commugrate/data/citeseer/.



(a) Before Update (b) After Update

Fig. 2. GUI for AVM-LB

Acknowledgments. The work was supported by the Ministry of Science and Technology of China, National Key Research and Development Program (Project Number 2016YFB1000703), the National Natural Science Foundation of China under No. 61732014 No. 61332006, No. 61472321, No. 61502390 and No. 61672432.

References

1. Batini, C., Scannapieco, M.: Data Quality: Concepts. Methodologies and Techniques. Springer Publishing Company, Incorporated (2010)
2. Fan, F., Li, Z., Wang, Y.: Cohesion based attribute value matching. In: 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), pp. 1–5, October 2017
3. Fan, F., Li, Z., Chen, Q., Chen, L.: Reasoning about attribute value equivalence in relational data. Inf. Syst. **75**, 1–12 (2018)
4. Naumann, F., Herschel, M.: An introduction to duplicate detection. Synth. Lect. Data Manag. **2**(1), 1–87 (2010)
5. Whang, S.E., Marmaros, D., Garcia-Molina, H.: Pay-as-you-go entity resolution. IEEE Trans. Knowl. Data Eng. **25**(5), 1111–1124 (2013)