Technical Perspective: Entity Matching with Quality and Error Guarantees

Benny Kimelfeld
Technion, Israel
bennyk@cs.technion.ac.il

Wim Martens
University of Bayreuth, Germany
wim.martens@uni-bayreuth.de

The challenge of *entity matching* is that of identifying when different data items (often referred to as *records* or *mentions*) refer to the same real-life entity. Popular instantiations of this problem include *deduplication*, where the items are database records that include duplicate representations of the same entity (e.g., duplicate profiles in a social network) [2], *record linkage*, where the items come from different data sources that mention overlapping sets of entities (e.g., the profiles of two social networks) [5], and *schema matching*, where the items are attributes of different database schemas that intersect on their domain of interest (e.g., the database schemas of different social networks) [6].

Common techniques for entity matching share various conceptual steps. First, *blocking* breaks the problem into considerably smaller subsets (blocks) of item pairs that have a reasonable chance to be matched, in order to reduce the quadratic number of needed comparisons. On each remaining pair to consider, a collection of *similarity functions* is applied to construct a vector of similarity scores. Next, a *classifier* transforms the vector into a decision: *match* or *non-match*. This classifier is typically built using supervised machine learning, where training is done over entity pairs labeled positively and negatively. Often, classification is complemented by a clustering algorithm if the matching is required to be transitive (i.e., if a profile matches a second profile, which matches a third profile, then the first must also match the third) [3].

There are other techniques for entity matching, including rule-based linking, and entity resolution via probabilistic inference. However, the field is generally short of fundamental guiding theory [4]. The paper "Entity Matching with Active Monotone Classification" [7] by Yufei Tao is a beautiful piece of work that proposes a principled approach to learn the aforementioned classification task over the vector of similarity scores, and more importantly, to reason about the theoretical bounds and the optimality of learning strategies.

The crux of the paper's development is to adopt an assumption that is very reasonable in the specific use case of the classifier: if *every* similarity function thinks that one pair is a better match than another, and if the latter is classified as a match, then the former should also be classified as a match. A classifier that features this behavior is called *monotone*, and the paper studies the learnability of monotone classifiers.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 2008 ACM 0001-0782/08/0X00 ...$5.00.

It is of course possible that no monotone classifier exists that is perfectly correct, i.e., perfectly separates matches from non-matches. Therefore, the author focuses on tradeoffs between the number of errors a classifier makes and the number of pairs that need to be *probed* (checked if they are a match or not).

The main algorithm in the paper, *random probe with elimination* (RPE) has several properties that could make it quite appealing to practitioners. It just consists of six lines of code and is extremely simple. Nevertheless, the author shows that it has favorable theoretical guarantees: it ensures an asymptotically optimal tradeoff between the number of probes and and the number of misclassified matches. Furthermore, as the algorithm is based on random sampling, it is expected to scale quite well.

Yufei Tao's paper not only offers us a nice blend between theory and practice, it is also a nice blend between databases and machine learning, which fits perfectly in some of the main research perspectives for the Principles of Data Management field [1].

1. REFERENCES

[1] S. Abiteboul et al. Research directions for principles of data management (Dagstuhl perspectives workshop 16151). *Dagstuhl Manifestos*, 7(1):1–29, 2018.

[2] A. Arasu, C. Ré, and D. Suciu. Large-scale deduplication with constraints using dedupalog. In *ICDE*, pages 952–963. IEEE Computer Society, 2009.

[3] P. Christen. *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Data-Centric Systems and Applications. Springer, 2012.

[4] L. Getoor and A. Machanavajjhala. Entity resolution: Theory, practice & open challenges. *PVLDB*, 5(12):2018–2019, 2012.

[5] T. N. Herzog, F. J. Scheuren, and W. E. Winkler. *Data quality and record linkage techniques*. Springer, 2007.

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

[7] Y. Tao. Entity matching with active monotone classification. In *ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems (PODS)*, pages 49–62, 2018.