Entity Resolution: Theory, Practice & Open Challenges

Lise Getoor
University of Maryland, College Park
getoor@cs.umd.edu

Ashwin Machanavajjhala
Duke University
ashwin@cs.duke.edu

ABSTRACT

This tutorial brings together perspectives on ER from a variety of fields, including databases, machine learning, natural language processing and information retrieval, to provide, in one setting, a survey of a large body of work. We discuss both the practical aspects and theoretical underpinnings of ER. We describe existing solutions, current challenges, and open research problems.

1. INTRODUCTION

Entity resolution (ER), the problem of extracting, matching and resolving entity mentions in structured and unstructured data, is a long-standing challenge in database management, information retrieval, machine learning, natural language processing and statistics. Ironically, different subdisciplines refer to it by a variety of names, including record linkage, deduplication, co-reference resolution, reference reconciliation, object consolidation, identity uncertainty and database hardening. Accurate and fast ER has huge practical implications in a wide variety of commercial, scientific and security domains.

Despite the long history of work on ER there is still a surprising diversity of approaches – including rule based methods, pair-wise classification, clustering approaches, and richer forms of probabilistic inference – and a lack of guiding theory. Meanwhile, in the age of big data, the need for high quality entity resolution is only growing. We are inundated with more and more data that needs to be integrated, aligned and matched before further utility can be extracted.

This tutorial brings together perspectives on ER from a variety of fields, including databases, machine learning natural language processing and information retrieval, to provide, in one setting, a survey of a large body of work. We discuss both the practical aspects and theoretical underpinnings of ER. We describe existing solutions, current challenges, and open research problems.

Proceedings of the VLDB Endowment, Vol. 5, No. 12

2. OUTLINE

Despite its long history, with some of the earliest work going back to the 1950s, ER remains an active area of research. In fact, with the emergence of "big data", the problem has enjoyed a renaissance in recent years. We will begin by surveying some of the latest motivating problems for ER in domains like advertising, online shopping, knowledge management, and network science, the changing landscape for ER, and why the problem continues to be so important (and challenging!). The rest of the tutorial is divided into three parts – ER theory, which reviews models and algorithms, ER practice, which focuses on techniques for scaling ER, and ER challenges, where we outline active research areas.

2.1 ER Theory

We begin by introducing a simple abstraction for the entity resolution problem. We categorize ER based on the type of input – *single-entity ER*, where all mentions correspond to a single entity type, *relational ER*, where real world entities are linked (like in a social network), and *multi-entity ER* representing the most general problem with potentially linked mentions of different entity types (e.g., products, sellers and reviews).

We survey classical techniques for ER, which assume that there exists a distance function between pairs of mentions. These techniques can be broadly classified as pairwise ER, where the decision to match a pair of mentions is made independent of other mentions, and cluster-based ER, where equivalence classes of entities are constructed via clustering. Pairwise ER is well suited for the problem of aligning two databases of the same set of entities (e.g., lists of restaurants from two sites). We survey common algorithms for computing similarity functions between mentions, and rule-based and probabilistic methods for pairwise and cluster-based ER. We also discuss techniques for computing cluster representatives, a.k.a. canonical entities, from database and machine learning communities.

We conclude this section by discussing the state of the art *collective* probabilistic inference techniques for multi-entity ER. These techniques are becoming popular due to an abundance of redundant mentions of entities on the Web that are also linked, and techniques that only consider one entity type and that ignore links perform poorly. We describe approaches based on multi-relational clustering algorithms, probabilistic generative models, and probabilistic logical languages, e.g. Markov logic networks and probabilistic soft logic.

2.2 ER Practice

Naive ER algorithms that compare every pair of mentions is $O(n^2)$. We will review efficient indexing, blocking, and message passing techniques, that can reduce the complexity to near linear time. In addition, distributed computation can also significantly improve scalability of ER algorithms, and we will review recent work on distributed ER.

Another important practical aspect is the evaluation of ER results. A variety of measures have been proposed; we will present some of the popular ones, and discuss some of the important differences. We conclude this section with a brief overview of ER systems that have been developed in the academia and the industry.

2.3 ER Challenges

Finally, we highlight a few open research directions, including ER in dynamic time varying data, large scale identity management, privacy, query-driven ER, and active learning or crowd-sourcing based methods for ER.

3. BIOGRAPHICAL SKETCHES

Lise Getoor is an associate professor in the Computer Science Department at the University of Maryland, College Park. Her main research interests are machine learning and reasoning under uncertainty, particularly in the context of structured and sem-structured data. She has published numerous articles in machine learning, data mining, database, and artificial intelligence forums. She received her PhD from Stanford University in 2001.

Ashwin Machanavajjhala is an Assistant Professor of Computer Science at Duke University. His primary research interests lie in data privacy, systems for massive data, and statistical methods for data integration. Previously, he was a Senior Research Scientist at Yahoo! Research. His doctoral dissertation, from Cornell University, on defining and enforcing privacy was awarded the 2008 ACM SIGMOD Jim Gray Dissertation Award Honorable Mention.

4. REFERENCES

[1] A. Arasu, M. Goetz, and R.Kaushik. On active learning of record matching packages. In *SIGMOD*, 2010.

[2] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom. Swoosh: a generic approach to entity resolution. *VLDB Journal*, 18(1), 2009.

[3] I. Bhattacharya and L. Getoor. A latent dirichlet model for unsupervised entity resolution. In *SDM*, 2006.

[4] I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery in Data*, 1(1), 2007.

[5] M. Bilenko, B. Kamath, and R. J. Mooney. Adaptive blocking: Learning to scale up record linkage and clustering. In *ICDM*, 2006.

[6] S. Chaudhuri, K. Ganjam, V. Ganti, and R. Motwani. Robust and efficient fuzzy match for online data cleaning. In *SIGMOD*, 2003.

[7] S. Chaudhuri, V. Ganti, and R. Motwani. Robust identification of fuzzy duplicates. In *ICDE*, 2005.

[8] P. Christen. *Data Matching*. Springer, 2012.

[9] W. Cohen and P. Ravikumar. A hierarchical graphical model for record linkage. In *Proc. of UAI*, 2004.

[10] W. Cohen, P. Ravikumar, and S. E. Fienberg. A comparison of string distance metrics for name-matching tasks. In *Proc. of IJCAI*, 2003.

[11] X. Dong, A. Halevy, and J. Madhavan. Reference reconciliation in complex information spaces. In *SIGMOD*, 2005.

[12] I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Society*, 64(2283), 1969.

[13] L. Gravano, P. Ipeirotis, N. Koudas, and D. Srivastava. Text joins for data cleansing and integration in an rdbms. In *ICDE*, 2003.

[14] M. A. Hernandez and S. J. Stolfo. The merge/purge problem for large databases. In *SIGMOD*, 1995.

[15] D. V. Kalashnikov, S. Mehrotra, and Z. Chen. Exploiting relationships for domain-independent data cleaning. In *SDM*, 2005.

[16] H. Köpcke, A. Thor, and E. Rahm. Evaluation of entity resolution approaches on real-world match problems. *PVLDB*, 3(1):484–493, 2010.

[17] N. Koudas, S. Sarawagi, and D. Srivastava. Record linkage: Similarity measures and algorithms. Tutorial at SIGMOD, 2006.

[18] A. McCallum, K. Nigam, and L. H. Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In *KDD*, 2000.

[19] A. McCallum and B. Wellner. Conditional models of identity uncertainty with application to noun coreference. In *NIPS*, 2004.

[20] D. Menestrina, S. E. Whang, and H. Garcia-Molina. Evaluating entity resolution results. In *PVLDB*, 2010.

[21] M. Michelson and C. A. Knoblock. Learning blocking schemes for record linkage. In *AAAI*, 2006.

[22] A. E. Monge and C. P. Elkan. An efficient domain-independent algorithm for detecting approximately duplicate database records. In *SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, 1997.

[23] H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching. In *NIPS*, 2003.

[24] V. Rastogi, N. Dalvi, and M. Garofalakis. Large-scale collective entity matching. In *PVLDB*, 2012.

[25] E. S. Ristad and P. N. Yianilos. Learning string edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

[26] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *SIGKDD*, 2002.

[27] W. Shen, X. Li, and A. Doan. Constraint-based entity matching. In *AAAI*, 2005.

[28] P. Singla and P. Domingos. Multi-relational record linkage. In *KDD*, 2004.

[29] P. Singla and P. Domingos. Entity resolution with markov logic. In *ICDM*, 2006.

[30] S. E. Whang, D. Menestrina, G. Koutrika, M. Theobald, and H. Garcia-Molina. Entity resolution with iterative blocking. In *SIGMOD*, 2009.

[31] W. E. Winkler. Methods for record linkage and bayesian networks. Technical report, Statistical Research Division, U.S. Census Bureau, 2002.