Tutorial: Uncertain Entity Resolution

Re-evaluating Entity Resolution in the Big Data Era

Avigdor Gal
Technion – Israel Institute of Technology

ABSTRACT

Entity resolution is a fundamental problem in data integration dealing with the combination of data from different sources to a unified view of the data. Entity resolution is inherently an uncertain process because the decision to map a set of records to the same entity cannot be made with certainty unless these are identical in all of their attributes or have a common key. In the light of recent advancement in data accumulation, management, and analytics landscape (known as big data) the tutorial re-evaluates the entity resolution process and in particular looks at best ways to handle data veracity. The tutorial ties entity resolution with recent advances in probabilistic database research, focusing on sources of uncertainty in the entity resolution process.

Keywords: Entity Resolution, Big Data, Uncertainty Management, Data Integration

1. INTRODUCTION

Integration of data has been the focus of research for many years now. At the data level, entity resolution (also known as record deduplication [12]) aims at "cleaning" a database by identifying tuples that represent the same entity. At the metadata level, schema matching and mapping and ontology matching and alignment identify ontological relationships between structured data descriptions (such as attributes, classes, *etc.*). The need for data integration stems from the heterogeneity of data (arriving from multiple sources), the lack of sufficient semantics to fully understand the meaning of data, and errors that may stem from incorrect data insertion and modifications (*e.g.*, typos and eliminations). With a body of research that spans over multiple decades, data integration has a wealth of formal models of integration [9, 6, 7, 1], algorithmic solutions for efficient and effective integration [13, 11, 8], and a body of systems, benchmarks and competitions that allow comparative empirical analysis of integration solutions [3, 4].

The evolution of data accumulation, management, and analytics, has recently led to coining the term *big data*. It encompasses technological advancements such as Internet of things (accumulation), cloud computing (management), and data mining (analytics), packaging it all together while creating a new and challenging research agenda. In the light of these landscape changes we analyze the impact of big data on data integration, aka *big data integration*.

Big data is commonly characterized via a set of "V"s, out of which four became prominent. Big data is characterized by *volumes* of data to be gathered, managed and analyzed. Volumes of data are not foreign to data integration. For example, entity resolution is applied in industry to hundreds of millions of records (*e.g.*, the 2010 U.S. Decennial Census [14]). *Velocity* is of a bigger concern for data integration, since tasks were often considered to be performed off-line. Therefore, it may be sufficient to process a census of the magnitude of the US in 30 hours. However, once new applications kick-in and data integration moves towards quick and (hopefully not so) dirty online processing, velocity needs to be improved. Big data *variety* is, in fact, the bread and butter of data integration. A massive cohort of work in data integration aims at homogenizing heterogeneous data sources.

The fourth "V", *veracity* involves the truthfulness and reliability of the data to be integrated. This "V" is especially important in data integration, for example, whenever integration of news feeds from social media is needed. Since social media is an important source of big data, big data integration should be concerned with the veracity of data.

In this tutorial we shall focus on the latter aspect, that of data veracity, in entity resolution. We propose to position this aspect in the well-established probabilistic database theory, pushing forward a vision, on which uncertainty becomes a mandatory aspect of the process of entity resolution, rather than being treated as a noise that needs to be discarded. As part of this tutorial, the audience is exposed to the state-of-the-art in entity resolution as means to a) conduct further research in this field and b) embed management of uncertainty in data integration projects.

Section 2 outlines the various topics to be covered. Next, we detail the target audience and prerequisite knowledge requirements (Section 3), followed by a brief professional biography of the presenter (Section 4). This tutorial is the result of a Ph.D. proposal of Batya Kenig, a Ph.D. candidate at the Technion, and a followup of a recent Information Systems Journal manuscript [11].

2. DETAILED DESCRIPTION

Entity resolution is a fundamental problem in data integration dealing with the combination of data from different

 Articles from this volume were invited to present their results at the 40th International Conference on Very Large Data Bases, September 1st - 5th 2014, Hangzhou, China.
Proceedings of the VLDB Endowment, Vol. 7, No. 13

sources to a unified view of the data. It is often the case that the datasets to be integrated contain information on the same real-world entity. Therefore, in order to integrate two or more data sources it is necessary to recognize representations that refer to the same real-world entity. In scenarios where unique keys are available across the datasets to be integrated the problem may be solved by a database join. However, in most cases no such identifiers are available. This, combined with the fact that the data may contain misspellings, be incomplete or incorrect raises the need for more sophisticated techniques. These techniques can be broadly classified into deterministic rule-based [9], probabilistic [6] and learning based techniques [13]. Other techniques treat a record as one long field and use variations of string similarity metrics to determine which records are similar [5].

Entity resolution is inherently an uncertain process because the decision to map a set of records to the same entity cannot be made with certainty unless a common key exists. Making deterministic decisions at various stages of the process may lead to inaccurate results and loss of information. The main goal of this tutorial is to point to the sources of uncertainty in the entity resolution process, discuss which types of uncertainties have been handled in the literature and suggest new methods for coping with various types of uncertainties.

The tutorial starts with an overview of big data. Next, we shall describe the various steps of entity resolution, focusing on the stages where uncertainty is introduced. Then we tie together entity resolution uncertainty and probabilistic databases. We show three algorithms for entity resolution that use probability theory at different stages of the entity resolution process, based on the works of Ioannou *et al.* [10], Beskales *et al.* [2] and Kenig and Gal [11]. We conclude with a set of open research questions, tying it back to the big data setting.

3. TARGET AUDIENCE AND PREREQUISITE KNOWLEDGE

The tutorial is aimed at researchers and practitioners alike. The proposed framework for managing uncertainty in entity resolution can assist researchers, and especially young researchers (e.g., Ph.D. students) in establishing a clean line of research in this field. For practitioners, the tutorial serves as an opportunity to push the limits of data integration projects by acknowledging and managing explicitly uncertainty.

The research on which the tutorial is based is rooted in the database community and influenced by information retrieval and machine learning research. The tutorial requires basic understanding of databases. The theoretic foundation will be presented gently, using examples rather than formal definitions.

4. INSTRUCTOR: BRIEF PROFESSIONAL BIOGRAPHY

Avigdor Gal specializes in various aspects of data integration with more than 100 publications in journals (Journal of the ACM (JACM), ACM Transactions on Database Systems (TODS), IEEE Transactions on Knowledge and Data Engineering (TKDE), ACM Transactions on Internet Technology (TOIT), and the VLDB Journal), books (Schema Matching and Mapping), and conferences (ICDE, CIKM, ER, CoopIS, BPM). He is the author of the book "Uncertain Schema Matching", part of Synthesis Lectures on Data Management (March 2011) and a co-author of a recent paper in the Information Systems Journal, "MFIBlocks: An effective blocking algorithm for entity resolution", that uses data mining for entity resolution. Avigdor Gal is a recipient of the prestigious Yannai award for excellence in academic education.[1]

5. REFERENCES

[1] Z. Bellahsene, A. Bonifati, and E. Rahm, editors. *Schema Matching and Mapping*. Springer, 2011.

[2] G. Beskales, M. A. Soliman, I. F. Ilyas, and S. Ben-David. Modeling and querying possible repairs in duplicate detection. *Proc. VLDB Endow.*, 2(1):598–609, 2009.

[3] P. Christen. Febrl -: an open source data cleaning, deduplication and record linkage system with a graphical user interface. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1065–1068, New York, NY, USA, 2008. ACM.

[4] P. Christen. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE Transactions on Knowledge and Data Engineering*, 99(PrePrints), 2011.

[5] W. W. Cohen. Data integration using similarity joins and a word-based information representation language. *ACM Trans. Inf. Syst.*, 18:288–321, July 2000.

[6] I. P. Fellegi and A. B. Sunter. A Theory for Record Linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[7] A. Gal, A. Anaby-Tavor, A. Trombetta, and D. Montesi. A framework for modeling and evaluating automatic semantic reconciliation. *The VLDB Journal*, 14(1):50–67, 2005.

[8] A. Gal and T. Sagi. Tuning the ensemble selection process of schema matchers. *Information Systems*, 35(8):845–859, 2010.

[9] H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C. Saita. Declarative Data Cleaning: Language, Model and Algorithms. In *Proc. of Int'l Conf. on Very Large Databases (VLDB)*, 2001.

[10] E. Ioannou, W. Nejdl, C. Niederée, and Y. Velegrakis. On-the-fly entity-aware query processing in the presence of linkage. *PVLDB*, 3(1):429–438, 2010.

[11] B. Kenig and A. Gal. Mfiblocks: An eective blocking algorithm for entity resolution. *Information Systems*, 38(6):908–926, Sept. 2013.

[12] F. Naumann and M. Herschel. An introduction to duplicate detection. *Synthesis Lectures on Data Management*, 2(1):1–87, 2010.

[13] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 269–278, New York, NY, USA, 2002. ACM.

[14] W. Winkler, W. Yancey, and E. Porter. Fast record linkage of very large files in support of decennial and administrative records projects. Technical report, 2010.

[1] http://pard.technion.ac.il/moshe-yanai/