Tutorial for the Web Information System Engineering Conference (WISE) 2014 on:

Blocking Techniques for Web-Scale Entity Resolution

George Papadakis
Institute for the Management of Information
Systems - Athena Research Center
gpapadis@imis.athena-innovation.gr

Themis Palpanas
Paris Descartes University
themis@mi.parisdescartes.fr

Entity Resolution constitutes one of the cornerstone tasks for the integration of overlapping information sources. Due to its quadratic complexity, a bulk of research has focused on improving its efficiency so that it can be applied to Web Data collections, which are inherently voluminous and highly heterogeneous. The most common approach for this purpose is *blocking*, which clusters similar entities into blocks so that the pair-wise comparisons are restricted to the entities contained within each block.

In this tutorial, we elaborate on blocking techniques, starting from the early, schema-based ones that were crafted for database integration. We highlight the challenges posed by today's heterogeneous, noisy, voluminous Web Data and explain why they render inapplicable the early blocking methods. We continue with the presentation of the latest blocking methods that are crafted for Web-scale data. We also explain how their efficiency can be improved by meta-blocking and parallelization techniques.

We conclude with a hands-on session that demonstrates the relative performance of several, state-of-the-art techniques, and enables the participants of the tutorial to put in practice all the topics discussed in the theory.

Topic Overview

Entities lie at the core of Web Data, as a large part of their information pertains to profiles that describe real-world entities. Typically, these profiles are scattered across different entity collections, such as Freebase, DBPedia, Geonames, etc. Entity Resolution (ER) is the task of inter-linking these complementary data sources and of deduplicating their content. Its inherent quadratic complexity, however, calls for approximate techniques that sacrifice effectiveness in order to enhance time efficiency. Blocking is the most popular of these approaches.

Early blocking methods were crafted for databases and, thus, were based on the assumption that entity profiles adhere to specific schemata, thus containing noise only in their attribute values. This assumption, though, is unrealistic in the context of Web Data, where entity profiles are described by a multitude of heterogeneous schemata, containing noise in their attribute names, as well. Recent advances overcome the schema constraints of early blocking methods, proposing novel techniques that are inherently crafted for heterogeneous entity profiles. They also exhibit significantly higher efficiency, scaling to voluminous collections of Web Data. In this tutorial, we provide a comprehensive summary of these novel blocking methods and organize them in a way that explains their relative performance and functionality. We also provide tools and data for experimenting with the state-of-the-art approaches.

Tutorial Outline

Total Duration 3 hours

- Introduction to Entity Resolution – Preliminaries & Challenges (15 minutes)
- Introduction to Blocking – Categorization of Blocking methods (15 minutes)
- Early, Schema-based Blocking Methods (20 minutes)
- Blocking Techniques for heterogeneous, noisy Web Data (20 minutes)
- Block Processing Techniques (20 minutes)
- Break – 10 Minutes
- Meta-blocking methods (20 minutes)
- Experimental comparison of blocking methods (20 minutes)
- Parallelization methods (20 minutes)
- Demonstration using the implementations in https://sourceforge.net/projects/erframework (20 minutes)

Short Biographies of Tutors

George Papadakis is a Postdoctoral Researcher at the Institute for the Management of Information Systems, Athena Research Center. Before that he worked as researcher at the NCSR ``Demokritos'', the L3S Research Center and the Institute of Communications and Computer Systems. He holds a Diploma in Computer Engineering from the National Technical University of Athens and a PhD from the Leibniz University of Hanover on “Blocking Techniques for efficient Entity Resolution over large, highly heterogeneous Information Spaces”. In addition to entity resolution, his research focuses on web data mining and has received the best paper award from ACM Hypertext 2011.

Themis Palpanas is a professor of computer science at the Paris Descartes University, France. Before that he was a professor at the University of Trento, Italy, and he has worked as a researcher at the IBM T.J. Watson Research Center, the University of California at Riverside, Microsoft Research and IBM Almaden Research Center. He is the author of eight US patents, three of which are part of commercial products. He has received three best paper awards and was General Chair for VLDB 2013. Professor Palpanas has been working on the field of Entity Resolution for the last 5 years, publishing relevant methods to major journals (TKDE) and conferences (WSDM).

References

1. Steven Euijong Whang, David Menestrina, Georgia Koutrika, Martin Theobald, Hector Garcia-Molina: *Entity resolution with iterative blocking*. SIGMOD Conference 2009.
2. Omar Benjelloun, Hector Garcia-Molina, David Menestrina, Qi Su, Steven Euijong Whang, Jennifer Widom: *Swoosh: a generic approach to entity resolution*. VLDB J. 18(1): 255-276 (2009).
3. Steven Euijong Whang, David Marmaros, Hector Garcia-Molina: *Pay-As-You-Go Entity Resolution*. IEEE Trans. Knowl. Data Eng. 25(5): 1111-1124 (2013).
4. Timothy de Vries, Hui Ke, Sanjay Chawla, Peter Christen: *Robust Record Linkage Blocking Using Suffix Arrays and Bloom Filters*. TKDD 5(2): 9 (2011).
5. Peter Christen: *A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication*. IEEE Trans. Knowl. Data Eng. 24(9): 1537-1555 (2012).
6. Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, Vassilios S. Verykios: *Duplicate Record Detection: A Survey*. IEEE Trans. Knowl. Data Eng. 19(1): 1-16 (2007).
7. George Papadakis, Ekaterini Ioannou, Claudia Niederée, Peter Fankhauser: *Efficient entity resolution for large heterogeneous information spaces*. WSDM 2011: 535-544.
8. George Papadakis, Ekaterini Ioannou, Claudia Niederée, Themis Palpanas, Wolfgang Nejdl: *Beyond 100 million entities: large-scale blocking-based resolution for heterogeneous data*. WSDM 2012: 53-62.
9. George Papadakis, Ekaterini Ioannou, Themis Palpanas, Claudia Niederée, Wolfgang Nejdl: *A Blocking Framework for Entity Resolution in Highly Heterogeneous Information Spaces*. IEEE Trans. Knowl. Data Eng. 25(12): 2665-2682 (2013).
10. George Papadakis, Georgia Koutrika, Themis Palpanas, Wolfgang Nejdl: *Meta-Blocking: Taking Entity Resolution to the Next Level*, in IEEE Trans. Knowl. Data Eng. (to appear).
11. Lars Kolb, Andreas Thor, Erhard Rahm: *Dedoop: Efficient Deduplication with Hadoop*. 1878-1881.
12. Kolb, L., Thor, A., Rahm, E.: *Block-based Load Balancing for Entity Resolution with MapReduce*. In: CIKM, pp. 2397–2400 (2011).
13. Kim, H., Lee, D.: *Harra: fast iterative hashed record linkage for large-scale data collections*. In: EDBT, pp. 525–536 (2010).
14. Rastogi, V., Dalvi, N.N., Garofalakis, M. N.: *Large-Scale Collective Entity Matching*. PVLDB 4(4): 208-218 (2011).
15. Demartini, G., Difallah, D.E., Cudré-Mauroux, P.: *Large-scale linked data integration using probabilistic reasoning and crowdsourcing.* VLDB J. 22(5): 665-687 (2013).
16. Batya Kenig, Avigdor Gal: MFIBlocks: An effective blocking algorithm for entity resolution. Inf. Syst. (IS) 38(6):908-926 (2013).
17. Yongtao Ma, Thanh Tran: TYPiMatch: type-specific unsupervised learning of keys and key values for heterogeneous web data integration. WSDM 2013: 325-334.