Web-scale, Schema-Agnostic, End-to-End Entity Resolution

George Papadakis
University of Athens, Greece
gpapadis@di.uoa.gr

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

ABSTRACT

Entity Resolution lies at the core of data integration, with a bulk of research focusing on both its effectiveness and time efficiency. Initially, most relevant works were crafted for structured, relational data that are described by a schema of well-known quality and meaning. With the advent of Big Data, though, these early schema-based approaches became inapplicable, as the scope of Entity Resolution moved to Web Data collections, which abound in noisy, semi-structured, voluminous and highly heterogeneous information. To address these inherent challenges of Web Data, recent works on Entity Resolution adopt a novel, schema-agnostic functionality that emphasizes scalability and robustness to noise.

In this tutorial, we take a close look on this line of research, organizing the state-of-the-art in the field into a scalable, schema-agnostic end-to-end workflow that consists of 4 steps. The first two focus on improving time efficiency through blocking, while the last two steps are dedicated to effectiveness: *(i) Block Building* clusters similar entities into blocks so as to restrict the originally quadratic complexity to comparing just pairs of entities that are highly likely to be matching. *(ii) Block Processing* further cuts down on the computational cost by discarding pairwise comparisons that are repeated or lack sufficient evidence for producing duplicates. *(iii) Entity Matching* carries out all comparisons in the final set of blocks, creating a similarity graph with a node for every entity and a weighted edge for every pair of compared entities. *(iv) Entity Clustering* partitions the nodes of the similarity graph into *equivalence clusters* such that every cluster contains all resources that correspond to the same real-world object. Special care is taken to highlight recent works that take the efficiency of these steps to the next level through massive parallelization, which is typically based on the MapReduce paradigm. The tutorial concludes with a hands-on session that involves our publicly available reference toolbox for Entity Resolution. This will allow the participants to put in practice all the topics discussed in theory, examining the relative performance of the main state-of-the-art techniques over established benchmark datasets.

ACM Reference Format:
George Papadakis and Themis Palpanas. 2017. Web-scale, Schema-Agnostic, End-to-End Entity Resolution. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 TOPIC AND RELEVANCE

Entities constitute the core organizational unit of Web Data, with their profiles assembling valuable information about real-world objects. Thus, various data management applications like query answering [20] are based on their semantics and connections in order to improve their performance. Typically, though, these applications require the integration of profiles that pertain to the same real-world object, but are scattered across different entity collections, such as Freebase[1], DBPedia[2] and Geonames[3]. *Entity Resolution* is the task of inter-linking these complementary data sources and of deduplicating their content [6].

Entity Resolution is a relatively old problem that was mainly crafted for structured (relational) data, which were described by schemata of known semantics and quality [4]. This schema knowledge allowed experts to develop customized solutions that simultaneously maximized precision and recall for the data at hand. For example, manually-defined rules determined whether two person entities are certain or likely matches simply by checking the similarity of their e-mail or of their address' zip code, respectively. Such rules restrict the computational cost to the comparison of the most informative attribute values, yielding both high effectiveness and time efficiency.

However, the schema-based approaches are inapplicable to Web Data, which abound in semi-structured entity profiles with unprecedented levels of noise and heterogeneity and a loose schema binding of unclear semantics. In more detail, Entity Resolution over Web Data is challenged by the well-known three Vs [6, 9]:

(1) *Variety*, which is caused by the absence of a database-like schema and by the rich diversity of the domains they cover (there are ∼2,600 different vocabularies in the LOD cloud, but only 109 of them are shared by more than one entity collection[4]),
(2) *Volume*, which is due to the large number of entity collections and of entities inside every collection (the LOD cloud alone contains almost 10,000 entity collections with ∼150 billion triples describing more than 55 million entities[4]), and
(3) *Veracity*, which stems from various forms of inconsistencies, noise or errors in entity profiles, due to the limitations of the automatic extraction techniques or of the crowd-sourced contributions.

In this tutorial, we explicitly focus on web-data Entity Resolution, examining methods that have been proposed in the literature for tackling the above three Vs. We organize them in the 4-step workflow depicted in Figure 1.

The first step, called *Block Building*, intends to tame the inherently quadratic time complexity of Entity Resolution, $O(n^2)$: the

[1] `https://www.freebase.com`
[2] `http://dbpedia.org`
[3] `http://www.geonames.org`
[4] http://stats.lod2.eu

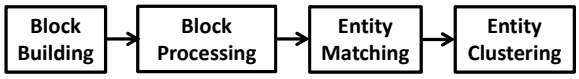

Figure 1: An End-to-End Workflow for Web-scale, Schema-Agnostic Entity Resolution.

naive, brute-force approach compares every entity profile with all others, a process that cannot scale to large entity collections. To boost time efficiency, Block Building reduces the executed comparisons to a significant extent at the cost of an approximate solution, i.e., by sacrificing recall to a minor extent. In essence, it clusters similar entities into blocks so that the pair-wise comparisons are restricted to the entities contained within each block [5]. For relational data, this process relied heavily on human intervention, requiring an expert to identify the best-performing attribute for extracting blocks from its values (e.g., the zip code for person entities). In contrast, Web Data call for automatic, generic methods that are able to build high quality blocks in a schema-agnostic way, i.e., by disregarding schema knowledge completely [6]. Our tutorial surveys the main relevant methods, explaining how they achieve high recall at the cost of very low precision.

To enhance block precision at a limited - if any - cost in recall, *Block Processing* is typically applied in the second step of the Entity Resolution workflow. Its goal is to refine the original blocks by efficiently removing comparisons that are repeated or involve non-matching entities [22]. Block Processing involves a series of methods and methodologies, like Meta-blocking [23, 30], which are by definition generic and schema-agnostic, thus applying naturally to Web Data. Our tutorial goes through the most prominent methods in the field, stressing how they can be combined for even better performance.

The next step in the Entity Resolution workflow is *Entity Matching*, which executes all comparisons contained in the refined set of blocks returned by Block Processing. Similar to Block Building, this process was simpler for relational data, since most techniques were based on value similarities, i.e., they employed stand-alone, schema-based pairwise comparisons. In contrast, Entity Matching for Web Data depends heavily on neighbor similarity, which is based on the relations between entities. Thus, it typically involves an iterative process that discovers duplicate entities gradually and propagates the latest matches to related entities that could benefit from them [18, 19, 31]. We review the main methods in the field, explaining the challenges posed by their iterative nature.

The end result of Entity Matching is a similarity graph, which contains a node for every entity and a weighted edge for every pair of entities that have been compared. To transform this intermediate model into the final outcome of Entity Resolution, *Entity Clustering* is applied. Its goal is to partition the graph nodes into equivalence clusters such that every cluster contains all duplicate entity profiles that actually correspond to the same real-world object [15]. Most relevant techniques are schema-agnostic by default, as they rely exclusively on the information contained in the similarity graph. Their time complexity, though, varies considerably, from linear to quadratic one. We examine the main clustering techniques in the literature from every type.

In order to foster further progress in the above areas, the tutorial also presents an open-source reference toolbox that implements the Entity Resolution workflow in Figure 1. This toolbox, named *JedAI* [27], is an open-source toolkit that involves numerous state-of-the-art methods for every step, including most of those discussed in theory. Thus, it enables users to build versatile workflows on-the-fly and can be readily used both for experimentation and for integration in Entity Resolution applications in the academic or the commercial domain, as it is distributed under the Apache License 2.0.

Finally, our tutorial takes special care to cover an important line of research in web-data Entity Resolution that gains more and more attention, namely *massive parallelization*. Recently, a bulk of work has been published in the field with the aim of exploiting the new parallelization paradigm, i.e., Map/Reduce [7]. We distinguish the relevant techniques into those parallelizing Block Building [11, 17], Block Processing [10] and Entity Matching in combination with Entity Clustering [3, 16, 29]. We also refer to systems that support parallelization, like LINDA [3] and Dedoop [16].

We note that contrary to most previous tutorials on the subject [8, 12–14], our tutorial surveys the state-of-the-art techniques for large-scale, schema-agnostic, end-to-end Entity Resolution. Its goal is to provide the participants with a deep understanding of the progress that has been made in the transition from solutions for homogeneous, structured data to solutions for heterogeneous, semi-structured Web Data. Furthermore, it equips participants with practical skills in applying Entity Resolution workflows and highlights the challenges that lie ahead in this active research area, discussing the latest works on progressive (online) [28, 34], crowdsourced [32] and query-driven [1] Entity Resolution.

Overall, our tutorial provides researchers with a complete coverage of the state-of-the-art Entity Resolution methods, as well as a discussion of a number of challenging open research problems that could well be the focus of their future research. Practitioners get a good overview of the benefits of main methods in the fields and learn how they can use them to improve the productivity of their businesses. They also learn to identify the methods or products that are more suitable for a particular task at hand, or better fit their general needs. The audience (and especially developers of information integration tools) additionally benefit from the hands-on session, learning how to integrate (parts of) the JedAI Toolkit into their applications. The developers also become acquainted with novel ideas that could well improve their existing products.

2 DURATION AND SESSIONS

The goal of our tutorial is to provide an overview of the state-of-the-art techniques for all steps of the End-to-End Entity Resolution workflow in Figure 1. To this end, every workflow step is analyzed in a different session. In total, our tutorial consists of 8 sessions, each lasting around 20 minutes, including 2 minutes for questions. Therefore, the intended duration of the tutorial is *half-day*. The content of the individual sessions is summarized below.

I **Introduction and motivation**
- Preliminaries on Entity Resolution
- Fundamental Assumptions, Principles and Definitions
- Entity Resolution Scope: Relational [4] vs Big Data [9]

II **Block Building**
- Taxonomy of Blocking Methods [21]

- Overview of Blocking for Relational Data [5]
- Blocking for Web Data [6]
- Experimental Analyses [5, 21]

III **Block Processing**
- Taxonomy of Pairwise Comparisons [26]
- Block Cleaning [22, 25]
- Comparison Cleaning [22, 25]
- Meta-blocking [23, 30]
- Experimental Analysis [26]

IV **Entity Matching**
- Overview of Methods for Relational Data [2]
- Methods for Web Data [18, 19, 31]

V **Entity Clustering** [15]
- Classification of Existing Methods
- Single-pass Clustering Algorithms
- Ricochet Family of Algorithms
- Other state-of-the-art Algorithms

VI **Hands-on Session** [27]
- The JedAI Open Source Library
- The JedAI Desktop Application
- The JedAI Workbench Tool - Demonstration with Benchmark Datasets

VII **Massive Parallelization**
- Parallel Block Building [11, 17]
- Parallel Block Processing [10]
- Parallel Entity Matching and Clustering [3, 16, 29]

VIII **Challenges and Final Remarks**
- Progressive Entity Resolution [28, 34]
- Crowd-sourced Entity Resolution [32]
- Query-Driven Entity Resolution [1]
- Conclusions

3 TARGET AUDIENCE

Our tutorial is example-driven, avoiding excessive technical details and proofs. As a result, there is no prerequisite knowledge, apart from a basic understanding of data management technology. This renders our tutorial suitable for a broad audience, covering not only students and researchers, but also practitioners and developers. In other words, it is intended for anyone with an interest in understanding the main techniques for scalable Entity Resolution over Web Data.

Besides the theoretical background in the state-of-the-art methods in the field, the participants will also gain practical skills through the hands-on session: they will get familiar with JedAI [27], our open-source reference toolbox, which incorporates the most prominent techniques in the area and can be readily used to tackle Entity Resolution problems.

4 PRESENTERS

George Papadakis is a Researcher at the Department of Informatics of the University of Athens, Greece, and an Internal Auditor of Information Systems at the main electricity company in Greece. Before that, he worked as researcher at the "Athena" Research Center, the NCSR "Demokritos", the L3S Research Center and the National Technical University of Athens (NTUA). He holds a Diploma in Computer Engineering from NTUA and a PhD from the Leibniz University of Hanover on "Blocking Techniques for efficient Entity Resolution over large, highly heterogeneous Information Spaces". His research interests pertain to Entity Resolution and Web Data Mining, in general. He has received the best paper award in ACM Hypertext 2011.

Themis Palpanas is a Senior Member of the Insitut Universitaire de France (IUF), and a professor of computer science at 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 nine US patents, three of which are part of commercial products. He has received three best paper awards, has been Associate Editor for PVLDB 2017 and General Chair for VLDB 2013, and is currently serving as Associate Editor for TKDE and PVLDB 2019, and Editor in Chief for BDR. Professor Palpanas has been working in the field of Entity Resolution for the last 8 years, publishing relevant papers in major journals and conferences.

5 PREVIOUS EDITIONS

An earlier version of this tutorial was presented by the same authors in the IEEE International Conference on Data Engineering (ICDE), 2016 [24]. Its duration was just 1.5 hours, as it focused exclusively on the first two steps of the end-to-end workflow in Figure 1, namely Block Building and Block Processing. It was attended by approximately 45 participants, in total.

The present tutorial is more extensive both in context and in scope: it contains entirely new content that covers the remaining workflow steps, i.e., Entity Matching and Clustering, while discussing in depth recent developments in the first two steps, like BLAST [30] and Semantic-aware Blocking [33]. Another difference is that we now consider techniques for progressive, crowdsourced and query-driven entity resolution. Finally, we present additional experimental results that demonstrate the relative performance of the main methods in every workflow step.

6 TUTORIAL MATERIAL

A website dedicated to our tutorial will be set-up two weeks before the presentation day. Initially, this website will give pointers and guidelines for the Entity Resolution toolkit that will be used during the hands-on session. All relevant code will be publicly released through the Apache License 2.0, which supports both academic and commercial uses. Finally, we will distribute all tutorial slides through the website one week before the presentation day.

REFERENCES

[1] Hotham Altwaijry, Dmitri V. Kalashnikov, and Sharad Mehrotra. 2013. Query-Driven Approach to Entity Resolution. *PVLDB* 6, 14 (2013), 1846–1857.
[2] Indrajit Bhattacharya and Lise Getoor. 2007. Collective entity resolution in relational data. *TKDD* 1, 1 (2007), 5.
[3] Christoph Böhm, Gerard de Melo, Felix Naumann, and Gerhard Weikum. 2012. LINDA: distributed web-of-data-scale entity matching. In *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012.* 2104–2108.
[4] Peter Christen. 2012. *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection.* Springer.
[5] Peter Christen. 2012. A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication. *IEEE Trans. Knowl. Data Eng.* 24, 9 (2012), 1537–1555.

[6] Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis. 2015. *Entity Resolution in the Web of Data.* Morgan & Claypool Publishers.

[7] Jeffrey Dean and Sanjay Ghemawat. 2008. MapReduce: simplified data processing on large clusters. *Commun. ACM* 51, 1 (2008), 107–113.

[8] Xin Luna Dong and Divesh Srivastava. 2013. Big Data Integration. *PVLDB* 6, 11 (2013), 1188–1189.

[9] Xin Luna Dong and Divesh Srivastava. 2015. *Big Data Integration.* Morgan & Claypool Publishers.

[10] Vasilis Efthymiou, George Papadakis, George Papastefanatos, Kostas Stefanidis, and Themis Palpanas. 2017. Parallel meta-blocking for scaling entity resolution over big heterogeneous data. *Inf. Syst.* 65 (2017), 137–157.

[11] Vasilis Efthymiou, Kostas Stefanidis, and Vassilis Christophides. 2015. Big data entity resolution: From highly to somehow similar entity descriptions in the Web. In *2015 IEEE International Conference on Big Data, Big Data 2015, Santa Clara, CA, USA, October 29 - November 1, 2015.* 401–410.

[12] Avigdor Gal. 2014. Tutorial: Uncertain Entity Resolution. *PVLDB* 7, 13 (2014), 1711–1712.

[13] Lise Getoor and Ashwin Machanavajjhala. 2012. Entity Resolution: Theory, Practice & Open Challenges. *PVLDB* 5, 12 (2012), 2018–2019.

[14] Lise Getoor and Ashwin Machanavajjhala. 2013. Entity resolution for big data. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013.* 1527.

[15] Oktie Hassanzadeh, Fei Chiang, Renée J. Miller, and Hyun Chul Lee. 2009. Framework for Evaluating Clustering Algorithms in Duplicate Detection. *PVLDB* 2, 1 (2009), 1282–1293.

[16] Lars Kolb, Andreas Thor, and Erhard Rahm. Dedoop: Efficient Deduplication with Hadoop. *PVLDB* 5, 12 (????), 1878–1881.

[17] Lars Kolb, Andreas Thor, and Erhard Rahm. 2012. Multi-pass sorted neighborhood blocking with MapReduce. *Computer Science - R&D* 27, 1 (2012), 45–63.

[18] Simon Lacoste-Julien, Konstantina Palla, Alex Davies, Gjergji Kasneci, Thore Graepel, and Zoubin Ghahramani. 2013. SIGMa: simple greedy matching for aligning large knowledge bases. In *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013.* 572–580.

[19] Juanzi Li, Jie Tang, Yi Li, and Qiong Luo. 2009. RiMOM: A Dynamic Multistrategy Ontology Alignment Framework. *IEEE Trans. Knowl. Data Eng.* 21, 8 (2009), 1218–1232.

[20] Davide Mottin, Matteo Lissandrini, Yannis Velegrakis, and Themis Palpanas. 2016. Exemplar queries: a new way of searching. *VLDB J.* 25, 6 (2016), 741–765. DOI: http://dx.doi.org/10.1007/s00778-016-0429-2

[21] George Papadakis, George Alexiou, George Papastefanatos, and Georgia Koutrika. 2015. Schema-agnostic vs Schema-based Configurations for Blocking Methods on Homogeneous Data. *PVLDB* 9, 4 (2015), 312–323.

[22] George Papadakis, Ekaterini Ioannou, Themis Palpanas, Claudia Niederée, and Wolfgang Nejdl. 2013. A Blocking Framework for Entity Resolution in Highly Heterogeneous Information Spaces. *IEEE Trans. Knowl. Data Eng.* 25, 12 (2013), 2665–2682.

[23] George Papadakis, Georgia Koutrika, Themis Palpanas, and Wolfgang Nejdl. 2014. Meta-Blocking: Taking Entity Resolutionto the Next Level. *IEEE Trans. Knowl. Data Eng.* 26, 8 (2014), 1946–1960.

[24] George Papadakis and Themis Palpanas. 2016. Blocking for large-scale Entity Resolution: Challenges, algorithms, and practical examples. In *32nd IEEE International Conference on Data Engineering, ICDE 2016, Helsinki, Finland, May 16-20, 2016.* 1436–1439.

[25] George Papadakis, George Papastefanatos, Themis Palpanas, and Manolis Koubarakis. 2016. Scaling Entity Resolution to Large, Heterogeneous Data with Enhanced Meta-blocking. In *Proceedings of the 19th International Conference on Extending Database Technology, EDBT 2016, Bordeaux, France, March 15-16, 2016, Bordeaux, France, March 15-16, 2016.* 221–232.

[26] George Papadakis, Jonathan Svirsky, Avigdor Gal, and Themis Palpanas. 2016. Comparative Analysis of Approximate Blocking Techniques for Entity Resolution. *PVLDB* 9, 9 (2016), 684–695.

[27] George Papadakis, Leonidas Tsekouras, Emmanouil Thanos, George Giannakopoulos, Themis Palpanas, and Manolis Koubarakis. 2017. JedAI: The Force Behind Entity Resolution. In *ESWC*. 161–166.

[28] Thorsten Papenbrock, Arvid Heise, and Felix Naumann. 2015. Progressive Duplicate Detection. *IEEE Trans. Knowl. Data Eng.* 27, 5 (2015), 1316–1329.

[29] Vibhor Rastogi, Nilesh N. Dalvi, and Minos N. Garofalakis. 2011. Large-Scale Collective Entity Matching. *PVLDB* 4, 4 (2011), 208–218.

[30] Giovanni Simonini, Sonia Bergamaschi, and H. V. Jagadish. 2016. BLAST: a Loosely Schema-aware Meta-blocking Approach for Entity Resolution. *PVLDB* 9, 12 (2016), 1173–1184.

[31] Fabian M. Suchanek, Serge Abiteboul, and Pierre Senellart. 2011. PARIS: Probabilistic Alignment of Relations, Instances, and Schema. *PVLDB* 5, 3 (2011), 157–168.

[32] Jiannan Wang, Tim Kraska, Michael J. Franklin, and Jianhua Feng. 2012. CrowdER: Crowdsourcing Entity Resolution. *PVLDB* 5, 11 (2012), 1483–1494.

[33] Qing Wang, Mingyuan Cui, and Huizhi Liang. 2016. Semantic-Aware Blocking for Entity Resolution. *IEEE Trans. Knowl. Data Eng.* 28, 1 (2016), 166–180.

[34] Steven Euijong Whang, David Marmaros, and Hector Garcia-Molina. 2013. Pay-As-You-Go Entity Resolution. *IEEE Trans. Knowl. Data Eng.* 25, 5 (2013), 1111–1124.