Blocking for Large-Scale Entity Resolution: Challenges, Algorithms, and Practical Examples

George Papadakis[1], Themis Palpanas[2]
[1]University of Athens, Greece gpapadis@di.uoa.gr
[2]Paris Descartes University, France themis@mi.parisdescartes.fr

***Abstract*—Entity Resolution constitutes one of the cornerstone tasks for the integration of overlapping information sources. Due to its quadratic complexity, a large amount of research has focused on improving its efficiency so that it scales 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 take a close look on blocking-based Entity Resolution, starting from the early blocking methods that were crafted for database integration. We highlight the challenges posed by contemporary heterogeneous, noisy, voluminous Web Data and explain why they render inapplicable these schema-based techniques. We continue with the presentation of blocking methods that have been developed for large-scale and heterogeneous information and are suitable for Web Data collections. We also explain how their efficiency can be further 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. The participants of the tutorial will put in practice all the topics discussed in the theory part, and will get familiar with a reference toolbox, which includes the most prominent techniques in the area and can be readily used to tackle Entity Resolution problems.

I. Introduction

Entities are becoming important for various data management applications, as they are increasingly being used (e.g., in query answering [1]) in order to exploit the semantics that they entail, or that can be derived through their connections to other entities. A large part of the information on entities pertains to profiles that describe real-world entities. Typically, these profiles 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.

The complexity of this cornerstone task is quadratic, since every entity profile has to be compared with all others. This means that it does not scale well to large entity collections. A bulk of research has focused on the problem of improving its efficiency – mainly through approximate techniques. These techniques may sacrifice recall to a small extent in order to significantly enhance time efficiency. The most popular method of this type 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-based Entity Resolution, surveying the main methods that have been proposed in this field. We start from the schema-based blocking methods that are crafted for the integration of structured data (i.e., databases). These methods rely on the assumption that entity profiles/records adhere to specific, a-priori known schemata, thus containing noise only in attribute values. We offer a comprehensive overview of these techniques, based on a taxonomy that facilitates their understanding. We explain, though, that their fundamental assumption is unrealistic in the context of Web Data, where entity profiles are described by a multitude of heterogeneous schemata, as they contain noise in attribute names, as well.

We then refer to recent advances that overcome the schema constraint, proposing novel techniques that are inherently crafted for heterogeneous entity profiles. They also exhibit significantly higher efficiency, scaling well to voluminous collections of noisy Web Data. We provide a comprehensive summary of the relevant techniques and organize them in a taxonomy that explains their relative functionality and performance.

Subsequently, we explain how the efficiency of all blocking methods can be further improved by block processing and meta-blocking techniques. These are generic techniques that aim to identify and discard either repeated comparisons, or comparisons between non-matching entities. We also present the major parallelization techniques that exploit the Map/Reduce framework for even higher efficiency. Special mention is given to the two main open challenges in the field, namely Crowdsourcing for blocking-based Entity Resolution and Incremental Blocking for Entity Resolution. We conclude with a hands-on session that demonstrates the relative performance of several, state-of-the-art techniques.

In the last five years, two ICDE tutorials have focused on the problem of Information Integration, which includes Entity Resolution, as well. The most recent one dealt with the integration of Big Data [2], while the other one examined data integration for Life Sciences [3]. None of them referred to (blocking-based) Entity Resolution in depth.

Instead, the goal of this tutorial is to provide an in-depth discussion of the abundant blocking techniques and the most recent advances in this area, which are the key techniques for

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

enabling Entity Resolution to operate effectively with large-scale and heterogeneous data.

II. Tutorial Outline

Our tutorial consists of 7 sections, each lasting between 10 and 15 minutes (1.5 hours, in total). At the end of every section, a question session of 2-minutes is provisioned. The content of the individual sections is summarized below.

A. Preliminaries & Challenges

We begin with an introduction to the main concepts of Entity Resolution, based on recent surveys [4] and books [5]. We explain the main challenges it involves and briefly describe the main approaches that address them. Then, we emphasize blocking as the most common solution for overcoming the challenge of scalability. We explain the benefits of blocking and refer to the qualitative measures that assess its performance [6]. We conclude with a taxonomy of the blocking methods that facilitates the understanding of the subsequent sections.

B. Blocking Methods for Databases

This section starts with an overview of the main blocking techniques that operate on top of structured data, such as Sorted Neighborhood [7], [8], StringMap [9], [10], CBlock [11] and MFIBlocks [12]. We explain why these methods are exclusively intended for Record Linkage and Deduplication of databases [5]. We distinguish them into supervised and unsupervised techniques and explain the relationship between the methods of each category. For each of the main methods, we briefly discuss the various enhancements that have been proposed in the literature. We also present an unsupervised approach for configuring their main parameter, i.e., their blocking keys, without any background knowledge [13]. We conclude with their pros and cons and present an assessment of their relative performance, based on the experimental analysis of a recent survey [6].

C. Blocking Methods for Web Data

In this session, we first motivate the need for blocking methods that are inherently crafted for Web Data. To this end, we highlight the new challenges these data collections pose compared to Databases. Based on these challenges, we stress the novel approaches that lie at the core of the blocking methods for semi-structured data. Then, we present an overview of the main relevant techniques, which include Token Blocking [14], TYPiMatch [15], URI Semantics Blocking [16] and Attribute Clustering [17]. We explain the relations between them and analyse their functionality through examples. We conclude with a comparative analysis of their performance over established datasets, stressing the pros and cons of each approach.

D. Block Processing Methods

Apart from the techniques that create blocks, our tutorial also considers methods that process an existing block collection in the optimal way. That is, in a way that discards most of the unnecessary comparisons, while retaining the comparisons between matching entities. In this category fall unsupervised techniques like Comparison Propagation [18], Meta-blocking [19], [20] and clustering-based techniques for specifying the optimal block sizes [21]. We additionally consider iterative techniques, like Iterative Blocking [22] and R-Swoosh [23], as well as supervised methods, such as Supervised Meta-blocking [24]. Again, we conclude with an experimental evaluation of the discussed techniques that highlights their pros and cons.

E. Parallelization

Another important line of research in blocking-based Entity Resolution is 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 [25]. We distinguish the relevant techniques into those parallelizing the creation of blocks and those parallelizing their processing. In the former category falls Dedoop [26], which parallelizes Standard Blocking and Sorted Neighborhood. The latter category includes the parallelization of Comparison Propagation [27], Parallel Meta-blocking [28] and Collective Entity Matching [29]. We also refer to systems that support parallelization, such as Linda [30].

F. Open Research Problems

Having summarized the main developments in blocking-based Entity Resolution, this section presents the open issues that currently lie at the focus of research. In particular, we examine two main topics: Crowdsourcing for Entity Resolution and Incremental Entity Resolution. For the former, we refer to recent advances, such as ZenCrowd [31], [32], and explain how they can be exploited in the context of (supervised) blocking methods. For the latter topic, we first explain that entity resolution is a continuous process, whose results have to be updated at the minimum cost, as the entity descriptions evolve, refreshing their data. Then, we briefly describe recent approaches [33], [34], [35], [36], and elucidate their relation to blocking methods.

G. Hands-on Session

The tutorial concludes with a hands-on session that is based on the publicly available application *Blocking Framework*[4], which has been developed by the authors. The application implements most of the techniques discussed in the tutorial and contains established benchmarks that allow for comparing their relative performance. The tutorial participants have the chance to download this application in their laptops, and learn how to use it during this last session.

The session begins with a presentation of the application's architecture that maps its structure to the sessions of the tutorial

[4]See `https://sourceforge.net/projects/erframework`.

so as to facilitate its use. Then, we explain how the individual techniques can be combined into comprehensive Entity Resolution workflows. We provide relevant guidelines, stressing that this process should be determined by the available resources and the performance requirements of the application at hand. Finally, we present the datasets that are available for experimentation and explain how they can be used for configuring a given workflow. At the end of the session, the participants will have the necessary knowledge and codebase to readily apply the state-of-the-art blocking techniques to their problems, putting in practice the topics discussed in theory.

III. Conclusions

Our tutorial surveys the state-of-the-art techniques for large-scale blocking-based 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 ones. Furthermore, it highlights the challenges that lie ahead in this active research area. To foster further progress in the open research problems, it also presents a reference toolbox that implements most of the discussed methods, and can be used both for experimentation and for integration in ER applications.

Our tutorial provides students and researchers with a complete coverage of the state-of-the-art blocking methods, enabling them to identify a number of challenging problems that could well be the focus of their future research. Practitioners get a good overview of the benefits of blocking methods 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. Developers of information integration tools benefit the most from the hands-on session, learning how to integrate (parts of) the Blocking Framework into their applications. They also become acquainted with novel ideas that could well improve their existing products.

References

[1] M. Lissandrini, D. Mottin, T. Palpanas, D. Papadimitriou, and Y. Velegrakis, "Unleashing the power of information graphs," *SIGMOD Record*, vol. 43, no. 4, pp. 21–26, 2014.

[2] X. L. Dong and D. Srivastava, "Big data integration," in *29th IEEE International Conference on Data Engineering, ICDE 2013, Brisbane, Australia, April 8-12, 2013*, 2013, pp. 1245–1248.

[3] S. C. Boulakia and U. Leser, "Next generation data integration for life sciences," in *Proceedings of the 27th International Conference on Data Engineering, ICDE 2011, April 11-16, 2011, Hannover, Germany*, 2011, pp. 1366–1369.

[4] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate record detection: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 1, pp. 1–16, 2007.

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

[6] ——, "A survey of indexing techniques for scalable record linkage and deduplication," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 9, pp. 1537–1555, 2012.

[7] M. A. Hernández and S. J. Stolfo, "The merge/purge problem for large databases," in *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, San Jose, California, May 22-25, 1995.*, 1995, pp. 127–138.

[8] ——, "Real-world data is dirty: Data cleansing and the merge/purge problem," *Data Min. Knowl. Discov.*, vol. 2, no. 1, pp. 9–37, 1998.

[9] L. Jin, C. Li, and S. Mehrotra, "Efficient record linkage in large data sets," in *Eighth International Conference on Database Systems for Advanced Applications (DASFAA '03), March 26-28, 2003, Kyoto, Japan*, 2003, pp. 137–146.

[10] C. Li, L. Jin, and S. Mehrotra, "Supporting efficient record linkage for large data sets using mapping techniques," *World Wide Web*, vol. 9, no. 4, pp. 557–584, 2006.

[11] A. D. Sarma, A. Jain, A. Machanavajjhala, and P. Bohannon, "An automatic blocking mechanism for large-scale de-duplication tasks," in *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*, 2012, pp. 1055–1064.

[12] B. Kenig and A. Gal, "Mfiblocks: An effective blocking algorithm for entity resolution," *Inf. Syst.*, vol. 38, no. 6, pp. 908–926, 2013.

[13] G. Papadakis, G. Alexiou, G. Papastefanatos, and G. Koutrika, "Schema-agnostic vs schema-based configurations for blocking methods on homogeneous data," *PVLDB*, vol. 9, no. 4, pp. 312–323, 2015.

[14] G. Papadakis, E. Ioannou, C. Niederée, and P. Fankhauser, "Efficient entity resolution for large heterogeneous information spaces," in *Proceedings of the Forth International Conference on Web Search and Web Data Mining, WSDM 2011, Hong Kong, China, February 9-12, 2011*, 2011, pp. 535–544.

[15] Y. Ma and T. Tran, "Typimatch: type-specific unsupervised learning of keys and key values for heterogeneous web data integration," in *Sixth ACM International Conference on Web Search and Data Mining, WSDM 2013, Rome, Italy, February 4-8, 2013*, 2013, pp. 325–334.

[16] G. Papadakis, E. Ioannou, C. Niederée, T. Palpanas, and W. Nejdl, "Beyond 100 million entities: large-scale blocking-based resolution for heterogeneous data," in *Proceedings of the Fifth International Conference on Web Search and Web Data Mining, WSDM 2012, Seattle, WA, USA, February 8-12, 2012*, 2012, pp. 53–62.

[17] G. Papadakis, E. Ioannou, T. Palpanas, C. Niederée, and W. Nejdl, "A blocking framework for entity resolution in highly heterogeneous information spaces," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 12, pp. 2665–2682, 2013.

[18] G. Papadakis, E. Ioannou, C. Niederée, T. Palpanas, and W. Nejdl, "Eliminating the redundancy in blocking-based entity resolution methods," in *Proceedings of the 2011 Joint International Conference on Digital Libraries, JCDL 2011, Ottawa, ON, Canada, June 13-17, 2011*, 2011, pp. 85–94.

[19] G. Papadakis, G. Koutrika, T. Palpanas, and W. Nejdl, "Meta-blocking: Taking entity resolutionto the next level," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 1946–1960, 2014.

[20] G. Papadakis, G. Papastefanatos, T. Palpanas, and M. Koubarakis, "Enhanced metablocking for scalable entity resolution over large, heterogeneous data," *EDBT*, 2016.

[21] J. Fisher, P. Christen, Q. Wang, and E. Rahm, "A clustering-based framework to control block sizes for entity resolution," in *KDD*, 2015, pp. 279–288.

[22] S. E. Whang, D. Menestrina, G. Koutrika, M. Theobald, and H. Garcia-Molina, "Entity resolution with iterative blocking," in *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2009, Providence, Rhode Island, USA, June 29 - July 2, 2009*, 2009, pp. 219–232.

[23] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. E. Whang, and J. Widom, "Swoosh: a generic approach to entity resolution," *VLDB J.*, vol. 18, no. 1, pp. 255–276, 2009.

[24] G. Papadakis, G. Papastefanatos, and G. Koutrika, "Supervised meta-blocking," *PVLDB*, vol. 7, no. 14, pp. 1929–1940, 2014.

[25] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[26] L. Kolb, A. Thor, and E. Rahm, "Dedoop: Efficient deduplication with hadoop," *PVLDB*, vol. 5, no. 12, pp. 1878–1881, 2012.

[27] ——, "Don't match twice: redundancy-free similarity computation with mapreduce," in *Proceedings of the Second Workshop on Data Analytics in the Cloud*. ACM, 2013, pp. 1–5.

[28] V. Efthymiou, G. Papadakis, G. Papastefanatos, K. Stefanidis, and T. Palpanas, "Parallel meta-blocking: Realizing scalable entity resolution over large, heterogeneous data," in *2015 IEEE International Conference on Big Data, Big Data 2015, Santa Clara, CA, USA, October 29-November 01, 2015*, 2015.

[29] V. Rastogi, N. N. Dalvi, and M. N. Garofalakis, "Large-scale collective entity matching," *PVLDB*, vol. 4, no. 4, pp. 208–218, 2011.

[30] C. Böhm, G. de Melo, F. Naumann, and G. Weikum, "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*, 2012, pp. 2104–2108.

[31] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux, "Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking," in *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, 2012, pp. 469–478.

[32] ——, "Large-scale linked data integration using probabilistic reasoning and crowdsourcing," *VLDB J.*, vol. 22, no. 5, pp. 665–687, 2013.

[33] Y. Altowim, D. V. Kalashnikov, and S. Mehrotra, "Progressive approach to relational entity resolution," *PVLDB*, vol. 7, no. 11, pp. 999–1010, 2014.

[34] T. Papenbrock, A. Heise, and F. Naumann, "Progressive duplicate detection," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1316–1329, 2015.

[35] A. Gruenheid, X. L. Dong, and D. Srivastava, "Incremental record linkage," *PVLDB*, vol. 7, no. 9, pp. 697–708, 2014.

[36] S. E. Whang and H. Garcia-Molina, "Incremental entity resolution on rules and data," *VLDB J.*, vol. 23, no. 1, pp. 77–102, 2014.