Zhou X, Hua W. Special section on entity resolution. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 35(4): 721–723 July 2020. DOI 10.1007/s11390-020-0005-5

Preface

It is our great pleasure to present the special section of Journal of Computer Science and Technology on "Entity Resolution (ER)". ER is the task of disambiguating various representations of the same real-world entity. It has been recognized as a key process for improving data quality in data integration of modern information systems. During recent decades, ER has developed beyond the traditional task of integrating database records, and has witnessed innovations in various domains including social media, multimedia, location-based services, etc. In addition, with the wide adoption and technological advancement of deep learning, much effort has been devoted to utilizing deep learning models for exploring the hidden behaviour of entities and thus enhancing the accuracy of linking. Moreover, ER has encountered a great challenge in the big data era to guarantee its efficiency, considering the prohibitively expensive operation of pairwise entity comparison, which requires new techniques to be developed and modern computing platforms to be utilized for efficient ER. Recognizing the growing impact of ER on information quality in organizations, and the new challenges and opportunities of ER in the big data era, this special section aims to report recent advances of new technologies, systems and frameworks that can support effective and efficient ER in different domains as well as new applications developed on top of ER.

We launched call for papers and sought original and high-quality research papers from all over the world. Nineteen submissions were received. A pre-review was conducted on all submissions and those of insufficient innovation or quality were immediately rejected. Two rounds of reviews were carried out for all the remaining submissions, with each paper assigned to at least two reviewers for a thorough review. Eventually we were able to accept 5 high quality submissions in terms of novelty, significance, clarity, and relevance. The accept rate is 26.3%. The accepted papers cover a broad range of topics, and we roughly classify them into three categories: domain-specific ER, efficiency issues in ER, and applications of ER.

Domain-Specific Entity Resolution

In addition to linking database records, ER has recently been applied in various domains such as Web data, heterogeneous information networks (HIN), social media, and so on. Deep learning models are widely adopted in these tasks, with different types of features being explored including entity attributes, textual context, graph structures, spatiotemporal behaviours, etc.

In "Enriching Context Information for Entity Linking with Web Data", the authors study the problem of disambiguating entities mentioned in natural language text. Existing solutions mainly focus on utilizing the local (i.e., contextual information in the text) and global (i.e., coherence among candidate entities) signals to generate the linking, which might be insufficient especially when dealing with informal and short texts. This paper proposes to enrich local and global information by getting extra context from the Web through search engines. Embedding-based methods and attention mechanisms are designed to generate high-quality Web context, which is then fed into a graph-based model for effective entity linking.

The paper "DEM: Deep Entity Matching Across Heterogeneous Information Networks" by Kong *et al.* targets at entity resolution for heterogeneous information networks. It proposes a new framework, DEM, to combine both entity attributes and network structure into a generic modelling, and utilizes multi-layer perceptron with a highway network to explore hidden relations and improve the accuracy of entity matching. Distributed network embedding is also introduced for enabling efficient computation of the matching in a vectorized manner.

Chen *et al.*, in the paper "User Account Linkage Across Multiple Platforms with Location Data", propose a novel model, ULMP, to support effective and efficient user account linking in the social media. They extend the

account linkage to multiple platforms, and explore users' spatiotemporal behaviours to retrieve the matched user accounts. More importantly, spatial and temporal index structures are designed to prune the search space and retrieve the top-k candidate accounts efficiently.

Efficiency Issues in Entity Resolution

ER is an expensive operation due to its inherently quadratic complexity. Many techniques have been designed to address the efficiency issues in ER and improve its scalability, including indexing, blocking, and filtering. In addition, modern computing platforms have also been exploited to support parallel processing of ER.

The paper "A Survey on Blocking Technology of Entity Resolution" by Li *et al.* presents a comprehensive survey on the mainstream blocking techniques developed in the recent decades. Blocking aims to cluster similar entities into blocks so that it is affordable to perform the pairwise comparison inside each block. It has been quite successful in reducing the computational cost of ER, with a slight sacrifice of accuracy. In this survey, the authors provide a summarization of classic blocking methods with emphasis on the analysis of different block construction and optimization techniques. Limitations of existing solutions are discussed, e.g., how to enable a more effective and more flexible utilization of schema information, machine learning, and deep learning models in entity resolution. Some promising directions of future work are also presented, such as real-time blocking, incremental blocking, and so on.

Applications of Entity Resolution

ER is an indispensable step in data integration and data management, which could benefit a vast amount of real-world applications including question answering, decision making, cross-domain entity profiling, personalized recommendation, etc., by bringing together information about entities from multiple sources.

The problem of linking-based cross-domain recommendation is investigated by Yi *et al.* in the paper "ATLRec: An Attentional Adversarial Transfer Learning Network for Cross-Domain Recommendation". The authors utilize the domain-sharable knowledge from auxiliary domains to alleviate the well-known data sparsity problem in existing recommender systems. The ATLRec model is proposed, based on adversarial transfer learning, to capture domain-sharable features through the interaction history of shared users, which are then combined with the domain-specific features to generate a more accurate recommendation.

As the leading editor and guest editor of this special section, we believe the selected papers present the new advances and leading topics in the field of entity resolution. We hope the special section will give readers some inspiration for their future work and stimulate further development in this area.

Finally, we would like to thank all the authors for submitting papers to this special section. We would also like to express our sincere gratitude to the reviewers of this special section who diligently assisted us in reviewing the papers. Without the reviewers' significant contributions, this special section would not have been published. We also thank the Editorial Director Ms Fengdi Shu for her encouragement, guidance, and help.

Leading Editor:

Xiaofang Zhou, Professor, IEEE Fellow, School of Information Technology and Electrical Engineering

The University of Queensland, Brisbane zxf@itee.uq.edu.au

Guest Editor:

Wen Hua, Assistant Professor, School of Information Technology and Electrical Engineering

The University of Queensland, Brisbane w.hua@uq.edu.au

Xiaofang Zhou is a professor of computer science at the University of Queensland (UQ), Brisbane, and the leader of Data Science Research Group at UQ, which includes the Data and Knowledge Engineering (DKE) Group. Prof. Zhou received his B.Sc. and M.Sc. degrees in computer science from Nanjing University, Nanjing, and Ph.D. degree in computer science from UQ. Before joining UQ in 1999, he worked as a researcher in Commonwealth Scientific and Industrial Research Organisation (CSIRO), leading its Spatial Information Systems group. His research focus is to find effective and efficient solutions for managing, integrating and analyzing very large amount of complex data for business, scientific and personal applications. He has been working in the area of spatiotemporal and multimedia databases, data mining, data quality, high performance query processing, big data analytics and machine learning, and co-authored over 300 research papers with many published in top journals and conferences such as SIGMOD, VLDB, ICDE, ACM Multimedia, AAAI, IJCAI, The VLDB Journal, ACM and IEEE Transactions. He was the Program Committee Chair of Australasian Database Conferences (ADC 2002 and 2003), International Conference on Web Information Systems Engineering (WISE 2004), Asia Pacific Web Conferences (APWeb 2003 and 2006), International Conference on Databases Systems for Advanced Applications (DASFAA 2009, DASFAA 2012 and DASFAA 2015), International Conference on Cooperative Information Systems (CoopIS 2012), IEEE International Conference on Data Engineering (ICDE 2013), ACM International Conference on Information and Knowledge Management (CIKM 2016), International Symposium on Spatial and Temporal Databases (SSTD 2017), and International Conference on Very Large Databases (VLDB 2020). He is a general co-chair of ACM Multimedia Conference 2015, IEEE International Conference on Data Management (MDM 2018), and China Big Data Technology Conference 2017. He has been on the program committees of numerous international conferences, often as a senior PC member, including SIGMOD, VLDB, ICDE, WWW, ACM Multimedia, ICDM, ICDCS and AAAI. He was the Convenor and Director of ARC Research Network in Enterprise Information Infrastructure in 2004–2011 (a major national research collaboration initiative in Australia), and the founding chair of ACM SIGSPATIAL Australian Chapter in 2010–2011. Currently he is an associate editor of IEEE Transactions on Cloud Computing, World Wide Web Journal, Distributed and Parallel Databases, Knowledge and Information Systems (since 2017), Springer's Encyclopedia of Database Systems, and Springer's Web Information System Engineering book series. He is the current Chair of IEEE Technical Committee on Data Engineering (TCDE, 2015–2018), and the Steering Committees of ICDE, DASFAA, WISE, APWeb and Australasian Database Conferences. In the past he was an associate editor of VLDB Journal (2008–2015), IEEE Data Engineering Bulletin, Information Processing Letters (2009–2015), IEEE Transactions on Knowledge and Data Engineering (2009–2015), and IEEE Transactions on Cloud Computing (2013–2019). He is a Fellow of IEEE.

Wen Hua is now working as a lecture (tenure track) and an Advance Queensland Research Fellow (AQRF) at School of Information Technology and Electrical Engineering, the University of Queensland (UQ), Brisbane. She received her Ph.D. and Bachelor's degrees in computer science from Renmin University of China, Beijing, in 2015 and 2010, respectively. From April 2016 to April 2017, she was appointed as a postdoctoral research fellow at UQ. She successfully won the AQRF Award in 2017 and obtained an ARC (Australian Research Council) Discovery Project as a chief investigator in 2019. She was the winner of the Best Paper Award in ICDE 2015. Her research mainly focuses on designing effective and efficient solutions for big data management, analysis, and knowledge discovery. She has been working in the areas of information extraction and retrieval, data mining, social media analysis, natural language processing, and spatiotemporal data analytics. She has published over 40 peer-reviewed papers in prestigious journals and international conferences including VLDB, SIGMOD, ICDE, VLDBJ, TKDE, SIGIR, CIKM, WSDM, IJCAI, WWWJ, etc. She has been actively engaged in professional services by serving as conference organizer, conference PC member for VLDB, ICDE, CIKM, ISWC, WISE, DASFAA, etc. and reviewer of more than 10 reputed journals such as VLDB Journal, TKDE, TMM, WWW Journal, JASIST, KAIS, and so on.