

Guest Editorial: Big Scholar Data Discovery and Collaboration

Yu-Ru Lin, Hanghang Tong, Jie Tang, and K. Selçuk Candan

◆

SCHOLARLY data collection, one of the most abundant resources for understanding human social activities and knowledge production processes dating back to 19th century, has expanded at a tremendous pace, with a variety of forms. Academics and researchers worldwide continue to produce large numbers of scholarly documents including papers, books, technical reports, etc., and associated data such as tutorials, proposals, and course materials. The ever-increasing diversity of disciplines and complexity of real-world problems, require researchers to seek new inspiration and collaboration outside of their own fields. Nowadays, besides traditional venues of collaboration such as conference meetings, the Internet provides a wide range of platforms for scholars to engage with other scholars. These new platforms (such as Google Scholar, ResearchGate, and Wi-ki-style virtual collaboration sites) enrich and document the ways scholars share academic resources, exchange opinions, follow each other's research, keep up with current research trends, and build their professional networks.

The rich combination of scholarly literature and the newly available information behavior data constitute a vast source of knowledge co-creation, search, and research traces. Being able to easily find and access useful information from such big scholarly data has raised a variety of research issues that involve data mining, data management, data curation and stewardship, system and platform design, data sharing policy, etc. For example, how do we better support different requirements of accessing, extracting and recommending scholarly articles and experts from the most up-to-date data? In this special issue, we aim to bring together researchers with diverse interdisciplinary backgrounds interested in scholarly big data challenges.

The first part of this special issue includes two articles that exemplify research opportunities and challenges in scholarly big data. Their contributions cover three themes: scalable algorithms for extracting scholarly-specific information content, fine-grained tools for analyzing and anticipating scientific impact, and system design for supporting scholarly-specific information query.

In the paper "AlgorithmSeer: A System for Extracting and Searching for Algorithms in Scholarly Big Data" by Tuarob, Bhatia, Mitra, and Giles, the authors explored the question of automatically identifying "algorithms" from a very large scholarly corpus. Algorithms are an example of scholarly-specific information that is crucial in the computational science but automatically extracting such information to better support scholarly search has been challenging due to their arbitrary formats. The authors proposed hybrid machine learning approaches to tackle this question and presented a system that demonstrated the feasibility of their proposed approaches.

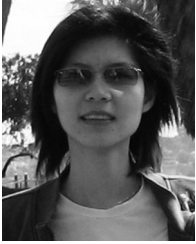
The paper "Can Scientific Impact Be Predicted?" by Dong, Johnson and Chawla brings the citation impact analysis to a more fine-grained level – how a scholar's citation impact will evolve in the future and which publications will contribute to it. The variability of individual trajectory and field conventions as well as their interplay made this problem particularly challenging. The authors reported a comprehensive empirical study of leveraging relevant factors for effective prediction, and presented an online tool demonstrating the utility of scholar-specific citation prediction.

ACKNOWLEDGMENTS

The authors would like to thank all the authors for considering this special issue as an outlet to publish their research results, and thank all the reviewers for their efforts and constructive feedback to the authors. They thank Dr. Huan Liu and Dr. Daqing He for their insightful discussions during the planning stage, and thank Yutao Zhang for his assistance in preparing the "Big Scholar Data Challenge" as part of this special issue. They also thank the staff of *TBD* and especially Kathy Santa Maria and Leigh Ann Testa for their help and guidance throughout the publication process. They are grateful for the support and advice that the editor-in-chief, Professor Qiang Yang, has given to them in preparation of this special issue.

-
- Y.-R. Lin is with the University of Pittsburgh. E-mail: yurulin@pitt.edu.
 - H. Tong and K.S. Candan are with Arizona State University.
E-mail: {hanghang.tong, candan}@asu.edu.
 - J. Tang is with Tsinghua University. E-mail: jietang@tsinghua.edu.cn.

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TBDDATA.2016.2562840

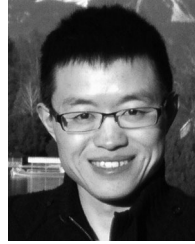


Yu-Ru Lin received the PhD degree in computer science with a concentration in arts, media and engineering from Arizona State University. She is an assistant professor at the School of Information Sciences, University of Pittsburgh. Her research interests include human mobility, social and political network dynamics, and computational social science. She has developed computational approaches for mining and visualizing large-scale, time-varying, heterogeneous, multi-relational, and semi-structured data. Her current research focuses on extracting system-level features from big data sets, including social media data and anonymized cellphone records, for studying human and social dynamics, particularly under exogenous events such as emergencies and media events. Her work has appeared in prestigious scientific venues including WWW, SIGKDD, SIGCHI, AAAI, InfoVis, *ACM Transactions on Knowledge Discovery from Data*, *ACM Transactions on Multimedia Computing, Communications, and Applications*, IEEEE, PLoS ONE, and EPJ Data Science. Her research vision is to use big data in the service of humanity, through developing new methodologies to collect, mine and utilize information to support collective sensemaking. Additional information can be found at: <http://www.yurulin.com/>.



Hanghang Tong received the MSc and PhD degrees from Carnegie Mellon University in 2008 and 2009, respectively, both majored in machine learning. He is currently an assistant professor at the School of Computing, Informatics, and Decision Systems Engineering (CIDSE), Arizona State University since August 2014. Before that, he was an assistant professor at the Computer Science Department, City College, City University of New York, a research staff member at the IBM T.J. Watson Research Center, and a Post-

doctoral fellow in Carnegie Mellon University. His research interest include large-scale data mining for graphs and multimedia. He has received several awards, including one test of time award (ICDM 10-Year highest impact paper award), four best paper awards and four best of conference. He has published more than 100 refereed articles and more than 20 patents. Additional information can be found at: <http://tonghanghang.org/>.



Jie Tang is an associate professor with the Department of Computer Science and Technology at Tsinghua University, and was also a visiting scholar at Cornell University. His interests include social network analysis, data mining, and machine learning. He has published more than 200 journal/conference papers and holds 20 patents. His papers have been cited by more than 6,000 times (Google Scholar). He served as a PC cochair of CIKM'16, WSDM'15, ASONAM'15, SocInfo12, KDD-CUP/Poster/Workshop/Local/Publication co-chair of KDD'11-15, and an associate editor-in-chief of *ACM Transactions on Knowledge Discovery from Data*, editors of *IEEE Transactions on Knowledge and Data Engineering/IEEE Transactions on Big Data*, and *ACM Transactions on Intelligent Systems and Technology*. He leads the project AMiner.org for academic social network analysis and mining, which has attracted more than 8 million independent IP accesses from 220 countries/regions in the world. He was honored with the Newton Advanced Scholarship Award, CCF Young Scientist Award, and NSFC Excellent Young Scholar. Additional information can be found at: <http://keg.cs.tsinghua.edu.cn/jietang/>.



K. Selçuk Candan is a professor of computer science and engineering at the Arizona State University. He has published more than 170 journal and peer-reviewed conference articles, one book, and 16 book chapters. He has nine patents. He served as an associate editor of one of the most respected database journals, the *Very Large Databases* journal. He is also in the editorial boards of the *ACM Transactions on Database Systems*, *IEEE Transactions on Knowledge and Data Engineering*, *IEEE Transactions on Cloud Computing*, *IEEE Transactions on Multimedia*, and the *Journal of Multimedia*. He has served in the organization and program committees of various conferences. In 2006, he served as an organization committee member for SIGMOD'06, the flagship database conference of the ACM. In 2008, he served as a PC chair for another leading, flagship conference of the ACM, this time focusing on multimedia research (MM'08). More recently, he served as a program committee group leader for ACM SIGMOD10. He also serves in the review board of the *Proceedings of the VLDB Endowment*. In 2011, he served as a general cochair for the ACM MM'11 conference. In 2012, he served as a general cochair for ACM SIGMOD'12. In 2015, he served as a general cochair for the IEEE International Conference on Cloud Engineering. He has successfully served as the PI or co-PI of numerous grants, including from the National Science Foundation, Air Force Office of Research, Army Research Office, Mellon Foundation, and HP Labs. He also served as a visiting research scientist at NEC Laboratories America for more than 10 years. He is a member of the Executive Committee of ACM SIGMOD and an ACM distinguished scientist. You can find more information about his research and an up-to-date resume at <http://aria.asu.edu/candan>.

▷ For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/publications/dlib.