Guest editorial

Guest editorial: managing bigger online data

Introduction

We are living in an era of massive data, whose volume, variety and velocity have not been seen before. *Harvard Business Review* reported that about 2.5 exabytes of data were created every day in 2012, and that this number would double every 40 months (McAfee and Brynjolfsson, 2012). Internet Live Stats (www.internetlivestats.com) states that the internet has over 1 billion websites at the time of writing, people send 500 million tweets each day, and Google receives over 3.5 billion searches in just one day. Furthermore, Facebook has 1.94 billion global monthly active users, and 1.74 billion are mobile users (www.statista.com).

Information about one topic is available in wide range of media and formats. For example, the AlphaGo's matches against two top human professional Go players, Lee Sedol and Ke Jie, in 2016 and 2017, were widely reported, discussed and presented on newspapers, TV programs, social platforms and various other online sites in many different languages and media. A complete understanding of the social impact of this event probably needs to consider information from all these sources.

Data are produced very quickly. For example, Facebook users produce more than 300 terabytes of log data every day and Taobao's 370 million members generate more than 20 terabytes of transaction data every day (Li and Cheng, 2012). At the same time, people increasingly demand immediate access to data. Users expect search engines to return relevant results within few seconds, drivers want live traffic information updated on their devices while they are driving and tweets are immediately pushed to subscribed followers for consumption. All of these are still relatively new and often are part of the so-called "big data" challenges.

Big data imposes challenges for academics as well. Researchers worldwide collect and generate massive volume of data stored in various forms of databases, and they continually produce large numbers of scholarly documents – including formal publications such as articles, books and technical reports, as well as informal documents such as tutorials, proposals, lab notes or course materials. To demonstrate the scale of academic contributions, PubMed has over 20 million medical-related articles with 10 million unique names and 70 million name mentions. In addition, an important shift in academics is the change of research paradigms in wide range of disciplines. Data-intensive and collaborative research has increasingly become the norm of many disciplines (Hey *et al.*, 2009). In such a paradigm, data are viewed as a critical component of the "infrastructure of science", which is important in forming "the basis for good scientific decisions" (Tenopir *et al.*, 2011).

However, as pointed by Borgman (2015) and many other researchers, the big data problem in academic disciplines does not always mean that the amount of data has to be at the petabyte or zetabyte level. There is a long tail distribution of research teams and the data they work with in their scholar activities. A large number of scholars only work with small amount of data. However, in this data-intensive research paradigm, they actually are facing even more big data-like problems. This is because as their research matures, data collection, analysis methods and storage and preservation facilitates may not be able to cope with the data that are larger and more diverse than before, as well as the increase in data itself. This "data exceeding current processing capacities" situation can be more accurately called the "bigger data" problem.

The Electronic Library
Vol. 35 No. 4, 2017
pp. 622-625
© Emerald Publishing Limited
0264-0473
DOI 10.1108/EL-08-2017-0187

"Bigger data" is a particularly important problem for the researchers in the library and information science (LIS) field. This is because LIS researchers can take a leadership role, not just participate in the management of bigger data. Borgman's statement of "having the right data is usually better than having more data" (Borgman, 2015, p. 4) clearly emphasizes the importance of managing bigger data. For long time, scholars and practitioners in LIS were the gatekeepers of human knowledge. The LIS field has developed strong technologies, methods, infrastructures and expertise on managing initially paper-based records and more recently digital materials. Many researchers, such as astronomers and biologists, may know how to process, analyze or mine data, but LIS experts, by working closely with these researchers, are able to explore the best approaches for managing knowledge discovery and organizing complex data. Thus, it is important for the LIS field to understand the latest development in bigger data related theories, technologies and practices for bigger data analysis, visualization, personalized service, privacy protection and complex network modelling, as well as bigger data preservation, sharing, reusing and other stewardship activities. All of these requirements bring great opportunities for the support and management of online bigger data.

Articles in this issue

Based on the understanding above, with the help of Professor Jiangping Chen and the whole editorial team of The Electronic Library, we organized this special issue on managing bigger online data. The specific topics that we called for included theory and methodology of bigger data management, online infrastructure and technologies for bigger data management and applications of bigger data management. With the active participation of the related communities, we selected 12 articles to include in this issue. These articles mainly cover four important topics in bigger online data management.

Social media mining and social network analysis

The paper "Emotion evolutions of sub-topics about popular events on microblogs" by Zhou and Zhang presents the first research study for emotion classification and evolution on extracted sub-topics of a popular event from user-generated content. Their subtopic identification uses a topic modelling approach, and their investigation of the emotion evolution is based on sentiment classification and time. Their results show that each sub-topic has its own specific primary emotions, which undergo different evolutions.

The paper "A belief-desire-intention model for blog users' negative emotional norm compliance: Decision-making in crises" by Wu *et al.* aims to understand blog users negative emotional norm compliance decision-making in crises (in short, the NNDC of blog users). They propose a belief–desire–intention model based on self-interests, expectations and emotions. Their model can explain the diffusion of negative emotions by blog users during crises, which provides a bridge between the social norm modelling and the research of blog users' behaviours in "real life" crises.

The paper, "Information diffusion on communication networks based on Big Data analysis" by Zhou *et al.* uses information diffusion models to investigate changes in the dissemination of information combing with data analysis. Their results validate the correctness of the proposed metrics and confirm that information spreads faster and wider with the development of information carriers.

Scientific text mining

Zeng *et al.*'s paper "The exploration of information extraction and analysis about science and technology policy in China" identifies the research gap of automatic content analysis of massive science and technology policies and proposes natural language processing technologies to extract and analyse important terms and sentences in science and technology policy documents. The goal is to provide a tested platform to facilitate users to quickly and effectively obtain valuable information from massive number of science and technology policy documents.

The paper "Semantically linking events for massive scientific literature research" by Zhang *et al.* presents a theoretical model, which represents a paper or a patent as a scientific research event. The elements of when, where, what, how and why are included in the model. Events and semantic relations among these elements are formulated into a semantic link network. Based on the network, event-centric information browsing, search and recommendation can be designed and developed.

Wang and Deng's paper "A paper-text perspective: Studies on the influence of feature granularity for Chinese short-text-classification in the Big Data era" is the first study to propose that Chinese characters, rather than terms or keywords, are more suitable as descriptive features in Chinese short-text-classification. They conducted their study by employing the categories discriminative capacity method on Chinese language fragments with different granularities and explore feasibility, rationality and effectiveness of Chinese characters in CSTC.

The paper "Exploring topics related to data mining on Wikipedia" by Wang and Zhang uses self-organizing maps and content analysis to explore the topics existing among Wikipedia articles that are related to data mining topics. The goal of this study is to gain insight into the general public's perceptions of data mining as a hot information technology topic. Through the categories identified in their study, they demonstrated that the general public is more interested in data mining organizations and applications of data mining in business than other topics. This helps to discover the gap between the general public and researchers.

User profiling and information behaviour study

Fang *et al.*'s paper "Measuring global research activities using geographic data of scholarly article visits" uses geographic data of the visits to scholarly articles to analyse the distribution of global research activities and to investigate the knowledge diffusion embodied in scientific papers. This study uses 23,798 articles published by 16 journals between 2007 and 2015, and the results show that visits were concentrated around major metropolitan areas and some high-tech clusters. In addition, new papers are initially visited by their publishing cities, and as time goes by there is diffusion to broader geographic areas.

The paper "Impact of device on search pattern transitions: A comparative study based on large-scale library OPAC log data" by Wu and Bi is among the earliest studies to focus on library OPAC users search behaviours from multiple devices, such as mobile phones, tablets, and desktops. Based on 9 GB transaction logs containing 16,140,509 records, they analyse the differences in search pattern transitions among different devices. Their results show that there are device impacts on users' search patterns, but such differences decrease when the iterations of query reformulation get higher.

The paper "Predicting users' demographic characteristics in a Chinese social media network" by Wang *et al.* aims to investigate multiple methods of constructing profiles for users on a Chinese social media platform. This investigation worked on 331,634 posts from 4,440 Sina Weibo users, and used a vector space model and topic models to construct users' profiles and predict their demographic characteristics including gender, age and geographic location. Theirs results show that latent semantic analysis performed better on the task of predicting gender and age, whereas the vector space model worked better in predicting geographic locations.

Data repositories and service

Yu's paper "The role of academic libraries in research data service (RDS) provision: Opportunities and challenges" examines RDS offered in academic libraries by combining existing literature and survey data from the association of research libraries (ARL) and the Association of College and Research Libraries. The goal is to illustrate the opportunities and challenges in providing RDS. Her results show that overall offerings of the library-led RDS in ARL research-intensive institutions have been increasing, which has two trends of increased engagement and expanded scope/level of services; however, discussions about RDS policy and infrastructure development are inadequate or largely non-existent.

The paper "Social science data repositories in data deluge: A case study of ICPSR's workflow and practices" by Jeng *et al.* investigates how the current practices in a data repository, the Interuniversity Consortium for Political and Social Research, map to the Open Archival Information System environment. Through conducting two focus groups and interviews, they examined the current actions (activities regarding their work responsibilities) and IT practices inside ICPSR for data curation. Their results show that OAIS model is robust and reliable in data curation and archive services, but at the same time, there are barriers and challenges for archiving and curating qualitative data at ICPSR.

Summary

Overall, managing bigger online data is an important topic, and these 12 articles represent some of the most important recent developments. We hope that researchers and practitioners in LIS can gain useful insights from these articles, which will inspire them to work further on this rapidly developing topic.

Xinning Su
Nanjing University, Nanjing, China
Chengzhi Zhang
Nanjing University of Science and Technology, Nanjing, China, and
Daqing He
School of Information Sciences, University of Pittsburgh,
School of Information Sciences, Pennsylvania, USA

References

Borgman, C.L. (2015), *Big Data, Little Data, No Data: Scholarship in the Networked World*, MIT Press, Boston, MA.

Hey, T., Tansley, S. and Tolle, K.M. (2009), *The Fourth Paradigm: Data-Intensive Scientific Discovery (Vol. 1)*, Microsoft Research, Redmond, WA.

Li, G. and Cheng, X. (2012), "Research status and scientific thinking of big data", *Bulletin of Chinese Academy of Sciences*, Vol. 27 No. 6, pp. 647-657.

McAfee, A. and Brynjolfsson, E. (2012), "Big data: the management revolution", *Harvard Business Review*, Vol. 90 No. 10, pp. 60-68.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A.U., Wu, L., Read, E. and Frame, M. (2011), "Data sharing by scientists: practices and perceptions", *PloS one*, Vol. 6 No. 6, p. e21101, available at: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0021101 (accessed 17 September 2017).