ISSN 1674 – 3393

A Peer-reviewed International Scholarly Journal

中国文献情报(季刊)

CHINESE JOURNAL OF LIBRARY AND INFORMATION SCIENCE

(QUARTERLY)

Volume 5 Number 3 2012
National Science Library, CAS

Chinese Journal of Library and Information Science (CJLIS)

(Quarterly)

Sponsored by the Chinese Academy of Sciences

Volume 5 Number 3, 2012

Published by: National Science Library, Chinese Academy of Sciences
No. 33, Beisihuan Xilu, Zhongguancun, Haidian District, Beijing 100190, P.R. China
Edited by: Editorial Office of *Chinese Journal of Library and Information Science* (CJLIS)
No. 33, Beisihuan Xilu, Zhongguancun, Haidian District, Beijing 100190, P.R. China
Printed by: Beijing KEXIN Printing Co. Ltd., Beijing 102208, P.R. China. Tel: 86-10-62903036. Fax: 86-10-62805493
Editor-in-Chief: Prof. Xiaolin Zhang
Typesetting: Beijing Charlesworth Software Dev. Co. Ltd. (Beijing Modern Palace Building)
No. 20 Dongsanhuan, RD(South) , Chaoyang District, Beijing 100022, P.R. China. Tel: 86-10-67791601. Fax: 86-10-67799806.
Distributed by: Editorial Office of *Chinese Journal of Library and Information Science* (CJLIS)
No. 33, Beisihuan Xilu, Zhongguancun, Haidian District, Beijing 100190, P.R. China
Subscription: RMB ¥ 200/Issue, RMB ¥ 800/Volume domestically per year; US $ 199/Volume outside of China (including air shipping)
Distributional Code (邮发代号) 82-563
Issue's Handling Editor: Jing CAO

ISSN 1674 – 3393
CN 11-5670/G2

Research Paper

A view on *big data* and its relation to Informetrics

Ronald ROUSSEAU[†1,2,3]

[1] Information and Library Science, Universiteit Antwerpen, B-2000 Antwerpen, Belgium
[2] Faculty of Engineering Technology, KHBO (Association KU Leuven), B-8400 Oostende, Belgium
[3] Department of Mathematics, KU Leuven, B-3000 Leuven (Heverlee), Belgium

Received Nov. 2, 2012
Revised Nov. 12, 2012
Accepted Nov. 15, 2012

Abstract

Purpose: *Big data* offer a huge challenge. Their very existence leads to the contradiction that the more data we have the less accessible they become, as the particular piece of information one is searching for may be buried among terabytes of other data. In this contribution we discuss the origin of *big data* and point to three challenges when *big data* arise: Data storage, data processing and generating insights.

Design/methodology/approach: Computer-related challenges can be expressed by the CAP theorem which states that it is only possible to simultaneously provide any two of the three following properties in distributed applications: Consistency (C), availability (A) and partition tolerance (P). As an aside we mention Amdahl's law and its application for scientific collaboration. We further discuss data mining in large databases and knowledge representation for handling the results of data mining exercises. We further offer a short informetric study of the field of *big data*, and point to the ethical dimension of the *big data* phenomenon.

Findings: There still are serious problems to overcome before the field of *big data* can deliver on its promises.

Implications and limitations: This contribution offers a personal view, focusing on the information science aspects, but much more can be said about software aspects.

Originality/value: We express the hope that the information scientists, including librarians, will be able to play their full role within the knowledge discovery, data mining and *big data* communities, leading to exciting developments, the reduction of scientific bottlenecks and really innovative applications.

Keywords *Big data*; CAP theorem; Knowledge representation; Data mining; Ethical concerns

CJLIS
Vol. 5 No. 3, 2012
pp 12–26
National Science Library, Chinese Academy of Sciences

† Correspondence: Ronald Rousseau (E-mail: ronald.rousseau@khbo.be). The author would like to thank Raf Guns, Qiaojun Hu and Brendan Rousseau for helpful comments. This article is an extended version of the presentation delivered during the 7th International Conference on Scientometrics and University Evaluation, October 26–28, 2012, Wuhan, China.

1 Introduction: What does the term *big data* mean?

We would like to start this contribution with a rhetorical question: Who were, in the course of history, the specialists in storing and retrieving large amounts of data? We all know the answer: Librarians and archivists of course. Yet, today in the digital area things have changed, and we ask: What does "large amounts of data" mean nowadays? What does the term *big data* refer to? The term "*big*" as used in the expression *big data* can mean several things: Its use depends on concrete circumstances. For some applications it means tens of terabytes ($10^{12} \approx 2^{40}$ bytes), while for large enterprises and huge scientific projects it may mean several petabytes ($10^{15} \approx 2^{50}$ bytes), even exabytes ($10^{18} \approx 2^{60}$ bytes).

According to Magoulas and Lorica[1] the term *big data* is used when the data size and the performance requirements become significant design and decision factors for implementing a data management and analysis system. The realm of *big data* is sometimes also referred to as non-relational database management systems (non-RDBMS). An important distinction between the realms of RDBMS and non-RDBMS is whether data are kept in a very structured manner (RDBMS) or not (non-RDBMS = *big data*). These two types are sometimes referred to as SQL and NoSQL databases, to differentiate the databases that use alternative solutions for storing large amounts of data from those that do not use SQL as their main query language.

There are three large domains in which *big data* lead to new challenges: Data storage, data processing and generating insights. The first two, namely, data storage and data processing, are studied within the computer sciences. As *big data* can be handled only difficultly using traditional relational databases or standard statistical packages, they require massively parallel software running on thousands of servers, such as in the so-called Google farms, located all over the world (including Hong Kong and Saint-Ghislain in Belgium) and consisting in total of more than 500,000 servers (and growing). Dean and Ghemawat[2] report that these servers process more than 20 petabytes of data per day.

In this article we will briefly touch on these computer related aspects, but it is the third one, namely, generating insights, which really belongs to the field of information science. Insights are obtained by using techniques such as text mining, audio and speech mining, social network analysis, advanced pattern recognition, data visualization and forecasting.

2 Where do these data come from?

Most of today's social activities (E-mails, Facebook, LinkedIn, Twitter) leave digital trails, leading to large, complex datasets. Besides social data which often

occur "by themselves", scientists collect and handle huge amounts of data within the framework of highly visible international projects. Examples are CERN's Large Hadron Collider (LHC), the Human Genome Project, ENCODE (a consortium and project that has built an encyclopedia of functional DNA elements)[3,4], the International HapMap Project, the 1,000 Genomes Project, the Human Variome Project, the Sloan Digital Sky Survey, etc., with many big projects collecting and re-analyzing climate data such as the 20th Century Reanalysis Project, and many more. Yet, already more than 13 years ago a warning about the coming tidal wave of data was published in *Nature*[5], while the main theme of *Nature* (2008, 455: 729, published in September 2008) was "science in the petabyte era". Hilbert and López discussed the enormous growth in information content made possible by the use of digital data[6].

Big data is not only an issue for scientists, but even more so for companies. They all collect data such as shopping data and try to make good use of them[7]. As this contribution is mainly aimed at scientists we just mention one example: TomTom, a Dutch firm of personal navigation systems, collected real travel times in Europe and North America and from these massive amounts of data constructed the TomTom Congestion Index (TTCI). This index compares travel times during non-congested periods (free flow) with travel times during peak hours[8,9]. The difference is expressed as a percentage increase in travel time. All types of roads are taken into account (local roads, arterials and highways) and raw data are actual global positioning system (GPS) based measurements. The most congested cities in Europe and North America are shown in Tables 1–2.

Table 1 Most congested European cities according to the TTCI – 2012[8]

Rank	City	Country	TTCI (%)	Morning peak (%)	Evening peak (%)
1	Warsaw	Poland	42	89	86
2	Marseille	France	41	79	81
3	Rome	Italy	34	76	66
4	Brussels	Belgium	34	82	86
5	Paris	France	32	72	63
6	Dublin	Ireland	30	70	62
7	Bradford-Leeds	UK	28	63	60
8	London	UK	27	48	48
9	Stockholm	Sweden	27	65	62
10	Hamburg	Germany	27	49	42

As *big data* are often produced by government institutes or projects supported by governments, governments, e.g. the Obama administration, also support *big data* studies[10].

Table 2 Most congested North-American cities according to the TTCI - 2012[9]

Rank	City	Country	TTCI (%)	Morning peak (%)	Evening peak (%)
1	Los Angeles	USA	33	56	77
2	Vancouver	Canada	30	51	65
3	Miami	USA	26	42	54
4	Seattle	USA	25	48	70
5	Tampa	USA	25	31	59
6	San Francisco	USA	25	51	62
7	Washington	USA	24	44	56
8	Houston	USA	23	41	65
9	Toronto	Canada	22	47	56
10	Ottawa	Canada	22	55	75

3 Essential properties and challenges: CAP theorem

As mentioned before, there are three large domains in which *big data* lead to new challenges: Data storage, data processing and generating insights. Data storage methods for *big data* have to deal with robustness issues and the scalability of the applied methods. However, one prefers a generic approach, this means that one system should be able to deal with all types of data. Data processing methods for *big data* should be able to read fast and provide fast updates and inserts. Again the applied methods must be extensible and the cost of maintenance may not be prohibitive.

Finally, insights must be generated in real-time and must provide information for debugging methods. Moreover, analysts must be able to perform not just standardized analyses, but also ad-hoc ones, on the fly.

A first step towards the solution of these problems is the use of distributed systems to achieve scalability and reliability.

3.1 CAP theorem, also known as Brewer's theorem

In the year 2000 Eric Brewer presented a talk at Berkeley (CA, USA) in which he discussed the CAP theorem (actually CAP conjecture at that time). The CAP conjecture states that it is only possible to simultaneously provide any two of the three following properties in distributed applications: Consistency (C), availability (A) and partition tolerance (P). Under certain conditions, including that such distributed systems are asynchronous and hence have no common clock, this conjecture became a theorem[11]. Here consistency roughly means that requests to the distributed systems must act as if they were executing on a single node, responding to operation ones at a time. Availability means that every request received by a non-failing node in the system must result in a response, while partition tolerance guarantees that the

system continues to operate despite arbitrary message loss or failures in parts of the system.

As users want highly available systems this leads in practice to systems for which the consistency requirement is relaxed so that updates are performed asynchronously. Readers must resolve potential data inconsistency[12]. Such systems often follow the so-called weak consistency model for which updates propagate over time and become eventually consistent.

3.2 Computer and software approaches

How are *big data* handled by real computer systems? MapReduce is a specific model for processing large datasets implemented by Google[2]. This model is named after the internal iterators in functional programming *map* and *reduce* on which it is based. A popular implementation of MapReduce libraries is Apache Hadoop. This is an open source software framework (sometimes referred to as an 'ecosystem') supporting data-intensive distributed (*big data*) applications.

4 An aside: Amdahl's law and its application to the theoretical study of collaboration

In the context of *big data* and the use of massively parallel computations one often refers to Amdahl's law. This law states that when a fraction of computations has to be done serially, then the maximum speedup obtainable from parallel processors is 1/s[13]. This law has been generalized by Kleinrock and Huang[14].

Egghe and Rousseau[15] applied the idea of serial and parallel processing on research collaboration, leading to considerations on the effectiveness of scientific collaboration. On the one hand, we have supposed that the more work is done in parallel the sooner the job is done, but on the other hand, the more time is spent in close collaboration (serially) the higher, probably, the quality gain. This leads to the problem, posed but not solved by Egghe and Rousseau[15], to find a compromise between *Too many cooks spoil the broth* and *Many hands make light work*.

These considerations lead us to the human side of parallelism. Organizing a dispersed group of scientists and technicians, supported by an administrative staff, is a difficult task. In a slightly different context, Rafols and Meyer[16] expressed this problem as a trade-off between the benefits of cognitive diversity and the costs for maintaining the cohesion of a team. Bringing highly capable and often very creative and original thinkers to give preference to the group's success and not the individual's need special leadership qualities. An account on these in the context of ENCODE is briefly described by Birney[3].

5 Insights: Knowledge discovery in databases (KDD)

Even before the era of *big data*, Fayyad et al.[17] wrote that because of the increase in computer power and the digital revolution more and more electronic data (not only text but also visual and auditive information) are collected while at the same time it has become easier to store it. Data are often kept (not deleted from the system) because one thinks that maybe, one day, these data can be of importance. In the sciences, data are sometimes collected with a lot of care and at a high price, for example, all data collected by satellites orbiting the earth. In a business environment, data are collected about daily sales, per point of sale, maybe even per customer. Large automated libraries collect data about the loan behavior of their clients. In a factory machines collect data about the production process. All these data contain the promise of more and better knowledge, ameliorated processes and avoiding making (the same) errors. The problem is, however, how can these promises become true? How can these massive data be manipulated to yield valuable, new information?

Raw data are not useful, at least not directly. Iron ore is not useful unless one knows how to make steel. Similarly, data are only a kind of raw material that may lead to possible useful knowledge. Traditionally data analysis has been a manual and slow process. Data analysts studied relatively large sets of data and laboriously uncovered patterns. Another complicating factor nowadays is the speed with which data grow. A manual approach is certainly too slow to keep up with this growth.

These circumstances have led to the emergence of the field known as knowledge discovery in databases (KDD) or data mining. The first KDD workshop has been held in 1989 and evolved into an international conference. Pattern recognition, statistics and AI-techniques are three important approaches used in KDD. Knowledge is often represented as rules, leading to rule-based systems, but this is only one possible form of knowledge representation. Such rules may, for instance, lead to assigning data to predefined categories or may define clusters describing the data[18]. Decision trees are another tool to represent knowledge found[19].

Discovering new facts in databases is described in different ways[17]. The term 'data mining' is often used by statisticians, database researchers and scientists investigating management information systems (MIS). Also colleagues associated with business schools often use this term. Others, such as Fayyad et al.[17], consider KDD as the general term and see data mining as one of 9 steps in a whole process as follows:

- Step 1: Searching (and finding!) what is already known in the area in which one hopes to find new knowledge and proposing a goal. Working in a totally unknown field is not only difficult but will moreover rarely lead to important

discoveries. Although the aim is to find new information, proposing a goal (while keeping an open mind) helps to focus the investigation.

- Step 2: Creating a 'target'-subset. It may happen that large parts of the original data set are totally unrelated to the set goal. Selecting an appropriate subset is then a useful thing to do which saves time and computer memory. Of course, such a data reduction is not always possible.
- Step 3: Cleaning data and preprocessing. In this step one tries to remove noise and obvious data errors (especially if this is possible in an automatic way). A decision has to be made about missing data, removing known trends, how to handle known changes (e.g. journals that are added or removed from the WoS) and known variants (e.g. name variants).
- Step 4: Data reduction. Now one tries to discover useful features, keeping the goal in mind. Dimension-reduction (multivariable statistics), transformations (Fourier transforms and wavelet transforms) can be applied to reduce the number of variables, leading to a better representation.
- Step 5: Making the goal of the data mining exercise more precise. What is the real goal of the exercise? Developing a new model? Building a new ontology or classification system? Should one use a regression model or apply a clustering technique? Is the main aim to find patterns in the data? It pays to be as precise as possible.
- Step 6: Choice of an appropriate algorithm. A choice has to be made about the algorithm one will apply. This algorithm will depend on the chosen goal and the type of data. Are data categorical or vectors with real coefficients? Are we interested in understanding a model, in predicting consumer behavior or future journal impact?
- Step 7: The actual data mining process. Everything that has been prepared is ready. The chosen algorithm is applied and results are shown using an appropriate representation, such as tree structures, networks, regression curves, clusters, time series or variance-covariance matrices.
- Step 8: Interpretation. Without interpretation results are worthless. In this step visualization techniques may be very helpful. Useless patterns are removed and other ones 'translated' for the user(s).
- Step 9: Use of the newly discovered knowledge. Now the new knowledge is ready to be applied and tested on its usefulness. Maybe existing evaluation or performance measures must be adapted. Maybe there is now a conflict between newly obtained and old knowledge, which must be resolved.

Note: If the application is in a scientific field, great care must be taken that all known facts are taken into account. Moreover, in science, rare events are often of utmost importance and may not be confused with errors or noise.

National Science Library, Chinese Academy of Sciences

6 Coming to the crux of the matter: Knowledge representation

Knowledge representation, here abbreviated as KR, is often considered as a part of the field of artificial intelligence. Yet, in our opinion, knowledge representation is also the core business of the information sciences. Indeed, searching for a book is not done by walking through the shelves and visually scanning all the books, but by using a representation of the library's content. In the old days, this was a card catalogue, then it was an OPAC available in the library and nowadays the contents of this OPAC is made available on the Internet so that the reader can check at home if the book is available in a particular library or not. Answering questions such as "How many books written by J.K. Rowling are among the library's holdings?" and "How many of these are at this moment available for loan?" can best be answered by using the representation (the catalogue, which nowadays usually includes a loan module). Library instruments such as lists of keywords, thesauri, taxonomies and ontologies are all forms of knowledge representation. This introduction makes clear the close relation between KR and the information sciences, even traditional librarianship. It leads us to a characterization of a KR.

6.1 The five characterizing roles of a knowledge representation

Davis et al.[20] discuss the five roles a knowledge representation must be able to play, claiming that these five properties characterize a bona fide KR.

- A KR is a surrogate.
- A KR is a (set of) ontological choices.
- A KR is a partial theory about intelligent reasoning.
- A KR is a medium for efficient calculations.
- A KR is a medium for human expression.

6.1.1 A KR is a surrogate

Physical objects, events, relations, persons cannot really be placed inside a computer. One uses a representation. This representation is a surrogate of the real thing. Symbols used in this representation form a model for reality (but are certainly not equal to it). Simulating reality and reasoning about it is then done by manipulating these surrogates. The example of the library catalogue shows that sometimes using a good KR is faster and more efficient than working in the real world.

6.1.2 A KR is a set of ontological choices

An efficient KR makes important aspects prominent and leaves out the unimportant ones. This decision process – what is of importance and what is not – entails an ontological choice.

6.1.3 A KR is a partial theory of intelligent reasoning

In a KR one must describe the exact relations between the objects included in the KR, and how to reason about them. Such a description is in fact a theory (often implicit) about intelligent reasoning (based on some form of logic). It must be clear which forms of reasoning are possible, allowed or meaningful and which are not.

6.1.4 A KR is a medium for efficient calculations (in the general sense)

Some problems are easy to represent but time-consuming to solve. Hence, in a KR the aspect time must be taken into account. Hence speed and efficiency of calculations allowed by the representation must be taken into account. This is especially true when dealing with large amounts of data.

6.1.5 A KR is a medium for human expression

Domain experts use knowledge representations to communicate with each other. Moreover, some KRs make it possible to explain, at least to some extent, difficult aspects of a theory or of reality to a lay public.

6.2 Some thoughts on applications of knowledge representations

Rich representations lead to reasoning about themselves. In this process one uses classical logic, but also newer forms of logic such as fuzzy logic and adaptive logic. Also the use of inheritance, frames and perspectives are of importance here[21]. Many problems can be represented using networks and graphs and can be converted to a search in such a network. Although the relation between 'data mining' and 'bibliometrics' and in particular 'citation analysis' is known for quite some time, it is only in recent years that citation analysis has become a tool among scientists performing data mining[22]. A network can be considered a form of knowledge representation, but it can also be considered as a structured data set, to be used for the detection of new knowledge. One promising approach is the use of link prediction techniques leading to the possible detection of missing links[23].

Detecting hot topics is another application within the field of *big data*[24]. Often mapping techniques help to visualize the data. One of the recent approaches is the use of overlays[25]. Recent mining techniques include sentiment analysis[26] and concept mining, where words are connected to the concepts they stand for[27,28].

The Web is a rich source for applications of mining techniques: Discovering relations, tools, services and documents[29]. Here we notice an overlap between the field of *big data* and webmetrics, hence informetrics.

7 Informetrics and a scientometric analysis of the *big data* field

Recall that the field of informetrics can be defined as the study of the quantitative aspects of information in any form, not just records or bibliographies, and in any social group, not just scientists. It includes the subfields of bibliometrics, scientometrics and webmetrics[30]. Clearly, all forms of data mining fall within this definition.

In a recent issue of Elsevier's *Research Trends*, the focus was on *big data*. Besides informetric studies, to which we come back in a moment, it covered the use of big datasets to inform funding and science policy decisions[31] and to help in developing science funding programs[32], an introduction to the so-called Fourth Paradigm, which is described as a way of new thinking stimulated by the availability of *big data*[33] and how *big data* analytics can be used at university level[34]. A practical illustration of the use of a *big data* approach was presented by Leetaru[35], by exploring and visualizing Wikipedia's view of world history. Finally, Moed[36] provided an example of combining multiple datasets, comparing citation and download data, discussing patents and questioning the validity of OECD input statistics.

Halevi and Moed[37] explored the use of the term *big data* in Scopus. They found 360 articles, the oldest ones much older than the invention of the term (with this special meaning). We performed a similar search in the WoS (TS= "Big data") on October 2, 2012, leading to 142 articles. We removed the oldest one (1974), and kept 141 published during the period 1993–2012. Halevi and Moed observed an over-exponential growth over the period (1970–2011), while we found a growth curve that could best be described by a cubic polynomial (R^2=0.963, with year 1992 = 0), which is illustrated in Fig.1.

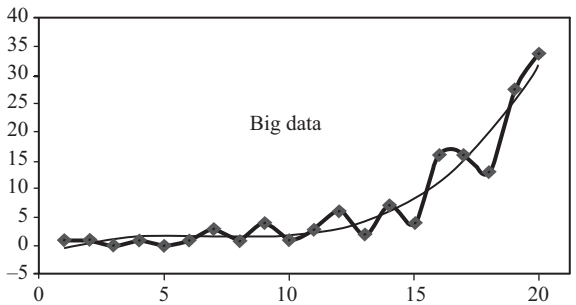

Fig. 1 *Big data* in the Web of Science.

During the period of 1993–2005 most articles were published in computer related fields, but also 5 of the 24 were in optics. Since 2006 one finds, besides computer-related fields, many articles in electronics and in multidisciplinary journals (bearing

testimony of the interest raised by the phenomenon – and the phrase –*big data*). The most used WoS categories are given in Table 3.

Table 3 *Big data* and most-used WoS categories in the period [2006–2012]

WoS category	# items
Computer science: Theory methods	19
Engineering: Electrical, electronic	19
Computer science: Information systems	16
Multidisciplinary sciences	15
Computer science: Artificial intelligence	12

Five articles on *big data* were classified in the category *Information Science and Library Science*, among which 2 published in the journal *Online* (WoS data). Halevi and Moed noticed that the most-occurring article type (in Scopus) was conference papers. The same is true in the WoS (Table 4), yet the ratio (articles)/(conference papers) is 0.8 in Scopus, while 0.87 in the WoS, indicating slightly different content structures in these two databases. We assume that the preponderance of proceedings papers has to do with the fact that *big data* are mainly studied in the field of computer sciences, where conference papers are often considered to be of more importance than journal articles.

Table 4 Article types in WoS (complete period)

Article type	# articles
Proceedings paper	54
Article	47
Editorial material	21
Letter	8
News item	7
Review	5

Finally, Table 5 shows the most active countries according to the two databases. Also here the two databases agree partially, but differ in the details.

Also Chinese journals, not included in the WoS, publish nowadays on *big data*. Good examples, involving a discussion of MapReduce are Qin et al.[38] and Zhang[39].

We conclude this informetric section by mentioning that also the field of *big data* has its own scientific journals. One of the latest additions is *GigaScience*, which describes itself as handling *big data* from the entire spectrum of the life sciences. This journal links articles directly to a database (GigaDB) hosting all underlying datasets and analysis tools. According to this journal's website the approach will overcome barriers to data sharing and make sure that research is reproducible. Indeed, irreproducibility of medical research results is nowadays a big issue[40], but becomes even more a focus point in the context of *big data*.

Table 5 Most active countries on the topic *big data*, according to Scopus[37] and the WoS

Scopus		WoS	
Country	# articles	Country	# articles
United States	101	United States	51
China	45	China	15
Germany	17	Germany	12
Japan	12	England	10
Italy	9	Spain	8
The Netherlands	9	Japan	6
Spain	9	Canada	5
UK	9	The Netherlands	5
Poland	7	Poland	5
Australia	6	Russia	5
Canada	6		
Russia	6		

8 Conclusion

Big data offer a huge challenge. Their very existence leads to the contradiction that the more data we have the less accessible they become, as the particular piece of information one is searching for may be buried among terabytes of other data. Yet, the methods and results of data mining processes offer rich potential for research in this area, providing fresh impetus to study *big data* in their social and institutional contexts. Analyzing such data in the sciences, maybe incorporating citation information may help us to understand how knowledge is obtained and how it changes over time[41]. Mathematical models may be of help in analyzing *big data* and when models fit well they may be used for prediction purposes and be useful for research policy decisions.

Yet, there still are serious problems to overcome. A trenchant critique concerning the *big data* field as it is nowadays came in the form of six statements intending to temper unbridled enthusiasm[42]. These six provocative statements are:

- *Big data* change the definition of knowledge;
- Claims to accuracy and objectivity are misleading;
- More data are not always better data;
- Taken out of context, *big data* lose their meaning;
- Just because it is accessible does not make it ethical; and
- (Limited) access to *big data* creates a new digital divide.

Surely boyd and Crawford[42] have brought the ethical dimension into the picture. Besides that they mention many theoretical and practical problems that still exist before the field of *big data* can deliver on its promises.

The TomTom Congestion Index (TTCI)[8,9] is the result of a study of traffic flows. Generally the field of *big data* must include a study of information flows leading to dynamically adapted insights.

As an information scientist in general with a special interest in citation network analysis I hope that our field will be able to play its full role within the, by definition interdisciplinary, knowledge discovery and data mining communities, leading to exciting developments, the reduction of scientific bottlenecks and really innovative applications.

References

1 Magoulas, R., & Lorica, B. Introduction to big data. Release 2.0. 2009: 11. Sebastopol, CA: O'Reilly Media. Retrieved on November 1, 2012, from http://radar.oreilly.com/r2/.
2 Dean, J., & Ghemawat, S. MapReduce: Simplified data processing on large clusters. Communications of the ACM, 2008, 51(1): 107–113.
3 Birney, E. Lessons for big-data projects. Nature, 2012, 489: 49–51.
4 Maher, B. The human encyclopedia. Nature, 2012, 489: 46–48.
5 Reichhardt, T. It's sink or swim as a tidal wave of data approaches. Nature, 1999, 399(6736): 517–520.
6 Hilbert, M., & López, P. The world's technological capacity to store, communicate, and compute information. Science, 2011, 332(6025): 60–65.
7 Brachman, R.J., Khabaza, T., & Kloesgen, W., et al. Mining business databases. Communications of the ACM, 1996, 39(11): 42–48.
8 TomTom. TomTom European Congestion Index, 2012. Retrieved on November 1, 2012, from http://www.tomtom.com/en_gb/congestionindex/
9 TomTom. TomTom North American Congestion Index, 2012. Retrieved on November 1, 2012, from http://www.tomtom.com/en_gb/congestionindex/
10 Gilbert, S., & Lynch, N.A. Brewer's conjecture and the feasibility of consistent, available, partition-tolerant Web services. ACM SIGACT News, 2002, 33(2): 51–59.
11 White House Office of Science and Technology Policy. Obama administration unveils "big data" initiative; announces $200 million in new R&D investments, 2012.
12 Shim, S.S.Y. The CAP theorem's growing impact. Computer, 2012, 45(2): 21–22.
13 Amdahl, G.M. Validity of the single processor approach to achieving large scale computing capabilities. In: AFIPS Conference Proceedings. Reston (VA): AFIPS Press. 1967, 30: 483–485.
14 Kleinrock, L., & Huang, J.H. On parallel processing systems: Amdahl's law generalized and some results on optimal design. IEEE Transactions on Software Engineering, 1992, 18(5): 434–447.
15 Egghe, L., & Rousseau, R. Amdahl's law and scientific collaboration. JISSI: The International Journal of Scientometrics and Informetrics, 1996, 2(1): 41–48.
16 Rafols, I., & Meyer, M. Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. Scientometrics, 2010, 82(2): 263–287.

National Science Library, Chinese Academy of Sciences

17 Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM, 1996, 39(11): 27–34.
18 Raghavan, V.V., Deogun, J.S. & Sever, H. Introduction (to special topic issue: knowledge discovery and data mining). Journal of the American Society for Information Science, 1998, 49(5): 397–402.
19 Vickery, B. Knowledge discovery from databases: An introductory overview. Journal of Documentation, 1997, 53: 107–122.
20 Davis, R., Shrobe, H.E., & Szolovits, P. What is a knowledge representation? AI Magazine, 1993, 14(1): 17–33.
21 Winston, P.H. Artificial Intelligence (Sec. Ed.). Reading: Addison-Wesley, 1984.
22 Guo, Z., Zhang, Z.F., & Zhu, S.H., et al. Knowledge discovery from citation networks. In: 9th IEEE International Conference on Data Mining (ICDM '09), 2009: 800–805.
23 Guns, R. Missing links. Doctoral dissertation, Antwerp University, 2012.
24 Porter, A.L., Newman, D., & Newman, N.C. Text mining to identify topical emergence: A case study on management of technology. In: Archambault, E., Gingras Y., & Larivière, V. eds. Proceedings of STI 2012. Montréal: Science-Metrix, 2012: 663–674.
25 Rafols, I., Porter, A.L., & Leydesdorff, L. Science overlay maps: A new tool for research policy and library management. Journal of the American Society for Information Science and Technology, 2010, 61(9): 1871–1887.
26 Prabowo, R., & Thelwall, M. Sentiment analysis: A combined approach. Journal of Informetrics, 2009, 3(2): 143–157.
27 Bichindaritz, I., & Akkineni, S. Concept mining for indexing medical literature. Engineering Applications of Artificial Intelligence, 2006, 19(4): 411–417.
28 Cox, A.M. Flickr: A case study of Web2.0. Aslib Proceedings, 2008, 60(5): 493–516.
29 Haverkamp, D.S., & Gauch, S. Intelligent information agents: Review and challenges for distributed information sources. Journal of the American Society for Information Science, 1998, 49(4): 304–311.
30 Tague-Sutcliffe, J. An introduction to informetrics. Information Processing and Management, 1992, 28: 1–3.
31 Lane, J. Big data. Science metrics and the black box of science policy. Research Trends, 2012, 30: 7–8.
32 Braveman, N.S. Guiding investments in research. Using data to develop science funding programs and policies. Research Trends, 2012, 30: 9–10.
33 Harris, R. ICSU and the challenges of big data in science. Research Trends, 2012, 30: 11–12.
34 Katz, D.S., & Allen, G. Computational & data science, infrastructure, & interdisciplinary research on university campuses. Research Trends, 2012, 30: 13–16.
35 Leetaru, K.H. A big data approach to the humanities, arts, and social sciences. Wikipedia's view of the world through supercomputing. Research Trends, 2012, 30: 17–30.
36 Moed, H. F. The use of big datasets in bibliometric research. Research Trends, 2012, 30: 31–33.
37 Halevi, G., & Moed, H. The evolution of big data as a research and scientific topic. Overview of the literature. Research Trends, 2012, 30: 3–6.

38 Qin, X.P., Wang, H.J., & Du, X.Y., et al. Big data analysis—Competition and symbiosis of RDBMS and MapReduce. Journal of Software, 2012, 23(1): 32–45.
39 Zhang, Z.Q. New paradigm for S&T intelligence studies. Journal of the China Society for Scientific and Technical Information = Qingbao Xuebao (in Chinese), 2012, 31(8): 788–797.
40 Naik, G. Scientists' elusive goal: Reproducing study results. Wall Street Journal Online. 2011. Retrieved on October 2, 2012, from: http://online.wsj.com/article/SB10001424052970203764804577059841672541590.html.
41 Liu, Y.X., & Rousseau, R. Interestingness and the essence of citation. Journal of Documentation (to appear in 2013).
42 boyd, D., & Crawford, K. Critical questions for big data. Provocations for a cultural, technological, and scholarly phenomenon. Information, Communication & Society, 2012, 15(5): 662–679.

News

Reference Citation Format

The format for citations in text and for bibliographic references follows GB/T 7714—2005. The citation should be ordered in number as it appears in the text of the submitted article.

- **For journal article**

  Sun, Y., Li, B., & Qu, J.F. Design and implementation of library intelligent IM reference robot. New Technology of Library and Information Service (in Chinese), 2011, 205: 88–92.

  Fernández, M., Kadiyska, Y., & Suciu, D., et al. SilkRoute: A framework for publishing relational data in XML. ACM Transactions on Database Systems, 2002, 27(4): 438–493.

- **For book**

  Cox, T.F., & Cox, M.A.A. Multidimensional scaling. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC, 2000.

  Campbell, N. (Ed.) Usability assessment of library-related websites: Methods and case studies. Chicago: Library & Information Technical Association, American Library Association, 2001.

  Hearst, M.A. User interfaces and visualization. In Ricardo, B.-Y., & Berthier, R.-N. (Eds.), Modern Information Retrieval. New York: ACM Press, 1999:257–323.

- **For proceedings**

  Åström, F. Visualizing library and information science concept spaces through keyword and citation based maps and clusters. In Bruce, H., Fidel, R., & Ingwersen, P., et al. (Eds.) Proceedings of 4th International Conference on Conceptions of Library and Information Science. Greenwood Village, CO: Libraries Unlimited, 2002:185–197.

- **For electronic journal article**

  Kurtz, M.J., Eichhorn, G., & Accomazzi, A., et al. The bibliometric properties of article readership information. Journal of the American Society for Information Science, 2004, 56(2): 111–128. Retrieved on May 3, 2005, from http://cfa-www.harvard.edu/~kurtz/jasist2.pdf.

  Järvelin, K., & Ingwersen, P. Information seeking research needs extension towards tasks and technology. Information Research, 2004, 10(1): Paper No. 212. Retrieved on May 3, 2005 from http://informationR.net/ir/10-1/paper212.html.

- **For electronic book**

  Crystal, A., & Ellington, B. Task analysis and human-computer interaction: Approaches, techniques, and levels of analysis. Proceedings of 10th American Conference on Information Systems, 2004: 1–9. New York. Retrieved on May 3, 2005 from http://www.ils.unc.edu/Nacrystal/AMC1504_crystal_ellington_final.pdf.

  Velterop, J. Open access publishing and scholarly societies: A guide. Retrieved on January 25, 2006, from http://www.soros.org/openaccess/pdf/open_access_publishing_amd_scholarly_societies.pdf.

- **For thesis or dissertation**

  Zhang, X.L. Information-seeking patterns and behaviors of selected undergraduate students in a Chinese university. Doctor Dissertation. New York: Columbia University, 1992.

CJLIS
Vol. 5 No. 3, 2012
pp 88–88
National Science Library, Chinese Academy of Sciences
http://www.chinalibraries.net

Submission Guidelines

◆ Aims

Chinese journal of Library and Information Science (CJLIS), being sponsored by the Chinese Academy of Sciences (CAS) and published quarterly by the National Science Library of CAS, is a scholarly journal in the field of library and information science (LIS). Its aim is to provide an international communication link between researchers, educators, administrators, and information professionals.

With the publication of the research results both from China and from other foreign countries, the Journal *CJLIS* strikes a balance between theory and practice. With its goal to provide an open forum for Chinese and international scholars in this field to exchange their research results, *CJLIS* also offers new possibilities in the advancement of Chinese library operations. The *CJLIS* tries to establish a platform for LIS students, researchers and library staff all over the world to engage in intellectual dialog and also to improve library services so as to promote even more quickened and substantial development of LIS in China.

◆ Scope

Striving toward academic excellence, innovation, and practicality, the *CJLIS* mainly includes research papers both on the theoretical as well as on the practical fronts in all aspects of the field. More specifically, it includes but not limited to informatics, library management, information technology application, knowledge organization system, knowledge management, archives, permanent preservation of library resources, LIS education, and so on.

◆ Refereeing Process

Articles and papers covering the topics or themes mentioned above, will be refereed through a double-blind peer review process.

◆ Editorial Advisory Board

The Editorial Board is composed of the nationally and internationally well-known scholars and researchers in the LIS field, the high quality of this Journal is thus reasonably assured.

◆ Manuscripts Categories

As the first English-language academic journal on LIS published in Mainland China, the *CJLIS* will take a proactive attitude to trace and report the prevailing hot issues in the field around the globe as well as the more serious scholarly communications. As such, the submitted manuscripts are classified into constant categories and unfixed categories. In the former category, research papers, library practice and progress reports are the essential components. In the latter, book reviews, biographical sketches, anecdotes, reminiscence of prominent librarians and brief communications will appear occasionally.

Research papers represent original research work or a comprehensive and in-depth analysis of a topic. More than 3,000 words are considered as a proper length for such manuscripts, with an abstract between 100 to 200 words.

Library practice covers the latest development and application in any segment of library field work and information service. The length of the manuscript is preferred to be more than 3,000 words, with an abstract between 100 to 200 words.

Progress reports reflect the projects result or research progress on the key topics of the library and information science. Submissions of articles to this section are expected to be comprehensive and analytical, which may deepen the understanding of the discussed issue and stimulate further researches on the topics, or give a new perspective on future technological applications. The manuscript length should be within 10,000 words, with an abstract between 100 to 200 words.

◆ Manuscripts Requirements

All papers can be submitted either in English or in Chinese (or both) with a double-line space. For the assurance that all the materials of the to-be-submitted are included, please check the following:

Title. Please give a brief biographical introduction to all contributing authors and their research background on a separated paper. For a better organization of the paper, please use the headings and subheadings.

Authors and affiliations. Please do not forget to write down the mailing address of each and every article contributors.

References. Be sure all the references used should be cited properly in both in-text and in bibliography. Particular attention should be paid to the proceedings. Do not forget adding the name(s) of editors of the compilation, as well as the name of the publishers. For the detailed information please request a copy of **Reference Citation Format**.

◆ Copyright

All submitted papers normally should not have been previously published nor be currently under consideration for publication elsewhere. For all the materials translated or obtained from other published resources, they should be properly acknowledged. All copyright problems should be cleared without any legal entanglements prior to the publication.

◆ Notes for Intending Submissions

A guide for authors and other relevant information, including submitting papers online, is available at the website of the Editorial Office of the *CJLIS* (http://www.chinalibraries.net). For any questions, you can e-mail the Office or directly to:

Prof. ZHANG Xiaolin
Editor-in-Chief of *CJLIS*
The *CJLIS* Editorial Office
National Science Library, Chinese Academy of Sciences
No.33 Beisihuan Xilu, Zhongguancun, Haidian District,
Beijing 100190, P.R. China
Tel: 86-10-82624454 or 86-10-82626611 ext. 6628
Fax: 86-10-82624454
E-mail: chinalibraries@mail.las.ac.cn
Website: http://www.chinalibraries.net

◆ Subscription

For single-copy subscription in China: RMB ¥ 200/Issue. For subscription outside of China, US $ 199/Volume yearly (including air shipping)

Chinese Journal of Library and Information Science (Quarterly)

Volume 5
Number 3
2012

Contents

ISSN 1674 – 3393
CN 11-5670/G2