ATLANTIS PRESS
Advances in Computer Science Research (ACSR), volume 73
7th International Conference on Education, Management, Information and Computer Science (ICEMC 2017)

Present and Future of Network Big Data

Xin Shi
1Beijing Language and Culture University, No.15 Xueyuan Road, Haidian District, Beijing, China
asamoashi@126.com

Keywords: Big Data; Network Big Data; Big Data Storage; Data Mining

Abstract. Network big data refer to the massive data generated by interaction and fusion of the Ternary human-machine-thing universe in the Cyberspace and available on the Internet. The Increase of their scale and complexity provided that of the capacity of hardware characterized by the Moore law, which brings the grand challenges to the architecture and the processing and computing Capacity of the contemporary IT systems, meanwhile set presents live opportunities on Deep mining and taking full advantage of the big value of network big data. Therefore, it is Pressing to research the disciplinary issues and discover the common laws of network big data, And further study the fundamental theory and basic approach to qualitatively or quantitatively Dealing with network big data. This paper analysis thezens caused by the complexity, Uncertainty and emergence of network big data, and summarizes major issues and research status of The awareness, representation, storage, management, mining, and social computing of network Big data, as well as network data platforms and applications. It is looking ahead to the development Trends of big data science, new modes and paradigm of data computing, new IT infrastructures, And data security and privacy, etc.

Introduction

In recent years, with the rapid development of the Internet, networking, cloud computing, network convergence and IT communication technology, the rapid growth of data has become a serious challenge faced by many industries and precious opportunities, so the information society has entered the era of big data (Big Data). The emergence of big data is not only changing the mode of operation of enterprises and the people's life and work, and even caused a fundamental change in the pattern of scientific research. At present, the rapid growth of network large data in size and complexity poses a challenge to the processing and computing power of existing IT architectures. According to a well-known consulting firm IDC released research report, the total network data in 2011 a total of 1.8ZB, is expected by 2020, the total will reach 35ZB.Network data to the academic circles also brings great challenges and opportunities. The direction of the emerging discipline of network data science and technology as the information science, social science, network science and system science and other related fields cross has gradually become a new hot topic in academic research. In recent years, the "Nature" and "Science" and other publications have been published. To investigate the effects of.2008 on the data of the year "Nature" published the monograph "Big Data", from the aspects of Internet technology, network economics, supercomputing, environmental science and biological medicine was introduced to the massive data challenge [1]. In 2011 "Science" launched on the data processing of special "Dealing with Data", discussed the data torrent (Data Deluge) brought opportunities [2]. In particular, it can effectively organize and use these data, people will get more opportunities to play a huge role in promoting the development of science and technology on society.

Challenges Posed by Network Big Data

Network large data is faced with challenges from many aspects. But from a research perspective, the fundamental challenge lies in its complexity, uncertainty and emptiness. The research on these three basic characteristics determines the development trend, research progress and application prospect of large network data.

The Complexity of Network Big Data. The development of information technology makes the data generated by the increasing channels, data types continue to increase. Accordingly, it is necessary to develop new data acquisition, storage and processing technologies. For example, the rise of social networks, making microblogging, SNS personal status information and other short text data gradually become the main information on the Internet media. Unlike traditional long texts, short texts are of great difficulty in traditional text mining (such as search, subject discovery, semantic, and emotional analysis) due to their short length, low context information and statistical information. Related studies include the use of external data sources (such as Wikipedia [3], search results [4], etc.) to expand the document, or use the internal similar document information to expand the expression of short text [5]. However, it is possible to introduce more noise, whether using external data or using internal data. On the other hand, the organic integration of different data types poses new challenges to traditional data processing methods. Such as the integration of geographical information and content in the study of social media [6], the combination of space-time information and content information [7] and so on.

The Uncertainty of Network Big Data. The inaccuracy of the original data and the granularity of the data acquisition and processing, the application requirements and the data integration and display make the data have different degrees of uncertainty in different dimensions and scales. The traditional focus on the accuracy of the data processing methods, it is difficult to deal with massive, high-dimensional, multi-type uncertainty data. Specifically, there are new ways to deal with the challenges of uncertainty in data collection, storage, modeling, querying, retrieval, and mining [8]. In recent years, the method of probability statistics has been gradually applied to the processing of uncertain data. On the one hand, the uncertainty of the data requires us to use an uncertain approach to deal with; the other hand, the development of computer hardware for such methods to provide the efficiency and efficiency of the possible. At present, the field of research is still shallow, in academia and industry there are still a lot of problems to be solved.

Storage and Management System of Network Big Data

Large scale network data processing data increased from grade TB to PB, EB, faced with how to reduce the cost of data storage, make full use of computing resources, improve the nonlinear iterative algorithm optimization problem many concurrent throughput, support the distributed system.

Distributed data storage. Row-Store and Column-Store are two typical database physical storage strategies. Row storage method is more traditional, it is in the disk to save each record in turn, more suitable for transaction operations; column storage method vertical division of the relationship table to column as a unit to store data, column storage also has data compression, Late materialization, Block Iteration, and so on [42]. Because data analysis tasks often use fewer fields, the column storage is more efficient. Data analysis tasks are more common in large data applications, so many systems, although they can not fully implement all the features of column storage, are more or less relevant concepts, including BigTable, HBase, etc. [43]. In [44], a series of mixed data storage structures (RCFile) are proposed to solve the problem of rapid loading of massive data, shorten query response time and efficient use of disk space (Figure 1). RCFile combines the advantages of row storage and column storage, reduces data load overhead by row grouping, and improves storage utilization through column data compression. The world's most widely used two distributed data analysis system Hive and Pig are integrated RCFile technology. RCFile has become a standard for data storage structures in distributed off-line data analysis systems.

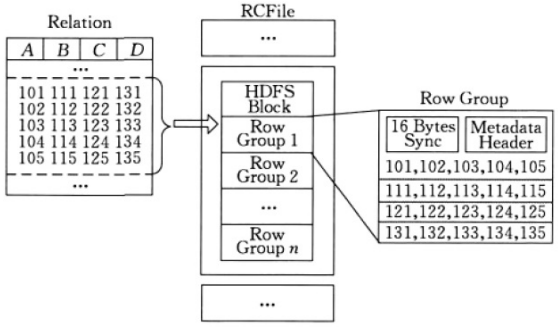

Figure 1. RCFile data storage structure

Distributed data storage is an important part of the application of network data. But the current research work still has some limitations. The amount of data to deal with massive data storage and processing of large scale, high processing speed and data type heterogeneity problem, need to support the development of a scalable, distributed processing depth above PB the data storage framework, also need to study to adapt to the data distribution storage structure optimization method to improve the network data storage and processing efficiency, reduce the cost of system construction, so as to realize network data efficiently and high availability of distributed storage.

Data Provenance Management. Data Provenance [9] contains the evolution of data evolution between different data sources and the internal data of the same data source. There are two basic methods, non-annotation methods, and annotation-based methods. The former uses the pattern mapping method to use the data processing function and its corresponding inverse function, but in a more complex case there may not be a reversible function between the sets, and the annotations must be used to describe the descent. In fact, the application of annotation-based methods is much higher than non-annotated methods.

Data systems can be used for a variety of data types, including relational data, XML data and uncertain data. Since the 1990s, the research of the data system has made great progress [10-11] and has been applied in many fields. In the face of large data on the network, the research of data system management needs to pay attention to the following aspects [12]: the traditional data management under the management of the Department of the Department of data there are still a lot of work to be considered, which study the origin and evolution of the data will be A large challenge; in the network environment uncertainty data is widespread, and has a variety of forms of expression. The evolution of data is accompanied by the evolution of data uncertainty, which can use the descent of data to track the source and evolution of data uncertainty. How to solve the problem of fusion of heterogeneous monastic standards. Large data applications will cover more of the original possible isolation of data sets, how to apply different standards of data to the lineage of information together is a key issue.

Summary

Cyberspace in the network of large data there is a huge scale, data association complexity, data status evolution and other significant features. Its size and complexity of the growth far beyond the Moore's Law in line with the growth of machine processing and computing power. Large network of data brings valuable opportunities, but also there are great challenges. This paper analyzes in detail the impact of these features on the depth analysis and value utilization of large network data. This paper reviews the network large data research system, network large data storage and management system, reviews the recent development of related fields, discusses the research direction and challenges of large network data, and looks forward to the future main research direction. In short, compared with the traditional research work, large differences in network data at all levels are very significant. Although there are already some exploratory research work, but in general, the network of large data research is still very young, there are still many problems to be solved.

References

[1] Big data. Nature, 2008, 455 (7209): 1-136
[2] Dealing with data. Science, 2011,331 (6018): 639-806
[3] Phan X H, Nguyen L M, Horiguchi S. Learning to classifyshort and sparse text & Web with hidden stories from largescale data collections, 2013 : 91-100
[4] Sahami M, Heilman T D. A web-based kernel function for measuring the similarity of short text snippets, 2015: 377-386
[5] Efron M, Organisciak P, Fenlon K. Improving retrieval of short texts through document expansion, 2012: 911-920
[6] Hong L, Ahmed A, Gurumurthy S, Smola A J, Tsioutsiouliklis K. Discovering geographical topics in the twitter stream, 2012: 769-778
[7] Pozdnoukhov A, Kaiser C. Space-time dynamics of topics instreaming text, 2011: 1-8
[8] Zhou Ao-Ying, Jin Che-Qing, Wang Guo-Ren, Li Jian-Zhong. A survey on the management of uncertain data. Chinese Journal of Computers, 2009, 32 (1): 1-16
[9] Gao Ming, Jin Che-Qing, Wang Xiao-Ling, Tian Xiu-Xia, Zhou Ao-Ying. A survey on management of data provenance. Chinese Journal of Computers, 2010,33 (3): 373-389
[10] Buneman P, Khanna S, Tan Wang-Chiew. Data provenance: Some basic issues, 2000: 87-93
[11] Tan W C. Provenance in databases: Past, current, and future. IEEE Data Engineering Bulletin, 2007, 30 (4): 3-12
[12] Gong Xue-Qing, Jin Che-Qing, Wang Xiao-Ling, Zhang Rong, Zhou Ao-Ying. Data-intensive science and engineering: Requirements and challenges. Chinese Journal of Computers, 2012,35 (8): 1563-1578