Indian Journal of Science and Technology, Vol 9(24), DOI: 10.17485/ijst/2016/v9i24/95993, June 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Hierarchical Group Data Management Scheme using Priority Information Big Data Environment

Nam-Kyu Park[1] and Yoon-Su Jeong[2*]

[1]Korea Institute of Science and Technology Information, Republic of Korea; Namkyu.park@kisti.re.kr
[2]Department of Information Communication Engineering, Mokwon University, Republic of Korea; bukmunro@mokwon.ac.kr

Abstract

Background/Objectives: With the advancement of mobile phone technology, services such as SNS and Facebook have become more popular and has dramatically increased the use of Big data. However, there are not many users who are satisfied the search results of their desired data. **Methods/Statistical Analysis**: This paper suggests a scheme that group-manages Big data by considering the similarity of data after first allocating priority to the data among a large volume of Big data. **Findings:** The suggested scheme pursues high accuracy and short processing time of the search results of Big data. In particular, the suggested scheme has faster processing velocity than existing scheme as it group-manages Big data by grouping the priority information according to the similarity allocated to data. **Application/Improvements:** The performance evaluation results indicated that the suggested scheme showed processing time 11.1% shorter and accuracy 8.3% better than the existing scheme on average.

Keywords: Big data, Data Management, Group Information, Mobile Phone, Priority

1. Introduction

As the number of Smartphone users rapidly increased following the recent advancement of IT technology, Big data service which is easily available is receiving great amount of attention[1,2]. In particular, the advancement of Big data technology, which is characterized by generation, collection, analysis and expression of large amount of data in diverse types, enabled more efficient operation of diversified contemporary society by making the prediction more accurate. It also made the provision, management and analysis of the customized information for each personalized member of the contemporary society possible, realizing the once impossible technology[3–6].

Big data is defined by the data amount in TB (Terabyte) unit and has a feature of increased volume of data as it requires long time for data collection and analysis[7,8]. However, rather than a simple increase of data volume, Big data is characterized by a complex change of the three largely divided elements: data volume, data velocity and data variety[9–11].

Big data suggests possibility of providing valuable information to society and human beings in all areas covering politics, society, economy, culture and scientific technology and its importance is ever growing[12]. However, as Big data is the set of countless information, a problem can arise in case private information is collected and managed during the collecting and analyzing process of data[13–15].

This paper suggests a group-management scheme of data through an inspection of similarity among information by allocating priority to data and hierarchically constructing link information among the priority information that was allocated so that users can quickly search what they want among the large volume of Big data information. In the suggested scheme, the group size of data is determined according to the priority of the hierarchically constructed data. As for the group data, data in the upper hierarchy include the data in the sub hierarchy and diverse attribute information which constructs data is subdivided in order to increase data accuracy. Moreover, the suggested scheme improves data accessibility compared

*Author for correspondence

to the existing scheme by linked-processing the priority information among data.

This paper consists as follows. Chapter 2 presents definition of Big data and reviews existing literature. Chapter 3 proposes a scheme of data group-management according to data priority. Chapter 4 conducts comparison evaluation between the suggested scheme and the existing scheme. Finally, Chapter 5 provides conclusion.

2. Related Works

2.1 Big Data

Big data has relatively larger volume and shorter generation cycle compared to the data that used to be generated in analogue environment in the past. Big data refers to large volume of data that not only contain numerical data, but also include text and video data [16-19]. Following the spread of PC, the Internet and mobile devices, data that are easily usable and storable in cyber space regardless of time and space are increasing by geometric progression[5,7]. The dissemination of Machine-to-Machine (M2M) which refers to the exchanges of information between human being and machine as well as machine and machine also provided ground for the exponential growth of digital information[3,11,14].

Video contents including UCC that users directly produce, text messages created in mobile phones and Social Network Service (SNS) show forms and quality that are different to the existing ones, in addition to the increased data velocity. In particular, text information on blogs or SNS makes the analysis of not only the personality of the individuals who wrote them, but also the connected relationship with counterparts possible. Moreover, CCTV is installed in road, public buildings and even elevators of condominiums where the videotaped image information is stored as data. In addition to the private sector, public sector is also creating data in bulk, including diverse social surveys such as the census, international data, medical insurances and pensions[2,4,8,15].

In general, Big data has a feature of 3V, which refers to the volume, velocity and variety of the type of the data. Diverse and vast amount of Big data are being used as important resource that determines the national competitiveness. However, a change of paradigm from an aspect of not only the quantity of data, but also their quality and variety is required considering their difference from the past[3,5].

Thanks to Big data that use technologies such as distributed processing system, prompt analysis on large amount of customer information became possible compared to the past. Firms can now analyze the company-related search words and comments generated in Twitter or on the Internet to understand the customer reaction on the company's products and service in real-time and implement instant response[9].

Since Big data make use of software or hardware of open-source type Hadoop or R, which is a package for analysis, analysis parallel processing technology and clouding computing, efficient system operation is possible without building high-cost data warehouse based on expensive storage and database[13,14].

2.2 Previous Research

Big data detection algorithms that have been studied so far were usually focused on the analysis of the text contents of microblog[15]. CELF algorithm is one of the most representative Big data detection algorithms that detects texts of microblog[6]. CELF algorithm selects subnet to detect every event if possible. However, CELF algorithm is unable to solve the mixed-integer optimization problem when selecting the optimal subnet.

A method of detecting major events online under resource restriction is suggested[7]. However, this scheme has a problem in that it should select and monitor small subnet to efficiently detect data such as small microblog. In[8], subnet of node is selected which has a maximal influence on data exploration where the effects account for the progression degree of the probability value from one node to another node. Even though this scheme does not mean the participation of events with a number of subnets with maximal influence, it is characteristic in that progression probability has great influence.

2. Hierarchical Data Management Method according to the Data Similarity

In this chapter, data are constructed such that they can be group-managed according to the data type, function and property by allocating priority to Big data information. The purpose is to improve the exploration accuracy and processing velocity of data. Moreover, by linking data that are constructed in groups according to their priority information, efficiency of data management can be improved.

3.1 Overview

Big data that exist in diverse types is one of the technologies that are receiving the largest attention along with the advancement of the Smartphone technology. It is important to accurately explore data and provide service in case of service such as SNS and Facebook. This paper aims at making data hierarchical by group-managing information with high similarity through a linked-process of priority that was allocated to the data and the attribute information so that users can accurately and quickly search their desired data in Big data environment.

The suggested scheme has an advantage of simply classifying and managing diverse and complicated data. In particular, it was assumed that the suggested scheme allocates data priority information beforehand when registering in server and provides service to the users so that the data used in the suggested scheme can be easily constructed and managed. Here, data reflects diverse attributes that fit the type and property of the data.

Figure 1 depicts the overall process of the suggested model. As is own in the picture, the suggested scheme constructs data in groups so that the distributed processing and storage management of the attribute information allocated priority information of the data can be possible. The size of the group can differ according to the correlation of linked information on data.

3.2 Hierarchical Data Grouping Construct

This section hierarchically groups data so that the data priority and attribute information can best correspond to the users' desired information as in Figure 2. This process is characteristic in that the data that are generated every second can be collected and managed in sample dataset.

Groups that construct the data with high similarity among them are linked to the data that have the high-

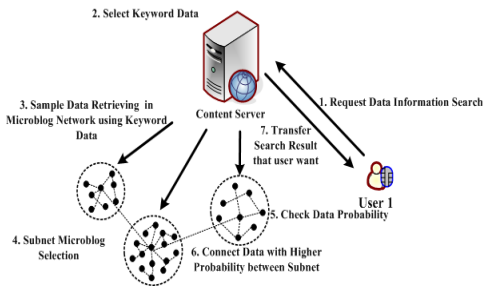

Figure 1. Overall process of proposed scheme.

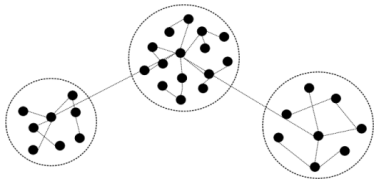

Figure 2. Group construction based probability.

est probability that represent the subgroup through the linked information. Here, the data that construct the subgroup are hierarchically constructed according to the probability value of the attribute information.

3.3 Subgroup Selection

To select subgroup, the suggested scheme first generates dataset through a sampling procedure of data of priority and attribute information that has high probability value among countless data of Big data. Here, it is assumed that data with probability value under a certain level, for example, less than a threshold ($P<0.3$), are filtered in the dataset to increase the accuracy and velocity. Once the filtering procedure is completed, the size of the sampling dataset for selecting a subgroup decreases. In the suggested scheme, dataset for detecting data is selected to choose the subgroup to extract priority and attribute information with high probability value.

3.4 Generation of Data for Construct the Subgroup

To generate data that construct the subgroup of Big data, the following five-phase process was performed.

- Phase 1: The number of users is assumed to be N. Among the N number of users U, the user who detects data and receives it from server is $U_i \subseteq U$.
- Phase 2: User U_i ($i \in [1, N]$) samples N data d to create dataset D_i. Dataset D_i is set as $i \in [1, N]$ and the sampled dataset is $D_i \subseteq L$. Here, L means total length of data that construct the dataset.
- Phase 3: In dataset D_i, data that are used in service are given with attribute value such as (fl_1, fl_2, …, fl_n). Here, fl_n means data property value and i means the element of set Z ($i \in Z$).
- Phase 4: Data probability included in dataset D_i is expressed as DP_i ($i \in [1, N]$). Data $\bar{d}$ allocated with data property value are given with binary probability information of 0 or 1 data according to the property

value. Here, data probability DP_i is expressed as Pr (P_i =1) or 1- DP_i = Pr (P_i = 0) and data probability is evaluated by $|D_i|/|D|$. $|\cdot|$ means the size of the dataset.

- Phase 5: Among the information of subgroup that is constructed by the data probability information, information $\overline{D}$ that has the highest data probability receives attribute information. According to the type, function and property of the relevant data, attribute set is generated and link information of data is created by applying to hash function H().

3.5 Data Linked-Process of Subgroup

When linked information on the data priority of subgroup among Big data information are generated, similarity among data is inspected and processed for data access as shown in Figure 3.

Figure 3 shows the construction of group-management of data by allocating the probability value according to the type, function and property of the data. Moreover, the efficiency of data management is improved by linked-processing data according to the priority information that were constructed in group.

Here, the suggested scheme calculates the similarity as probability for the segmented processing of data and extracts data with high data similarity to separately process the hierarchical data.

4. Evaluation

Performance evaluation was conducted on the suggested scheme to compare its processing time and accuracy with Shen scheme[10].

4.1 Environment Setup

The number of keywords for searching data in the suggested scheme has a set of {1, 3, 5, 10} and the number of data has a set of {500, 1000, 2500, 5000, 7500, 10000}. The number of attributes of data was set at 5.

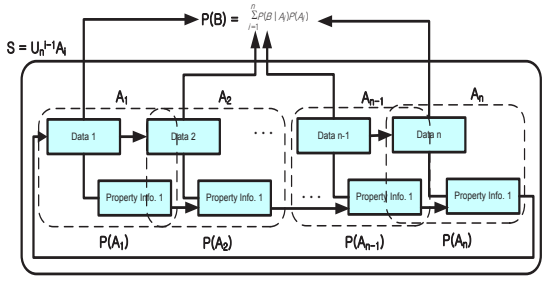

Figure 3. Procedure of data linked-process of subgroup.

4.2 Performance Analysis

4.2.1 Processing Time

Figure 4 shows the processing time for extracting data that belong to subgroup by linking data through probability information about the data priority and attribute information after hierarchically grouping data in order to obtain the data required by the users among Big data. Test results in Figure 4 indicates 11.1% decrease of the processing time in case of the suggested scheme on average compared to Shen scheme[10]. These results can be attributed to the use of probability link information according to the data priority and attribute information in the suggested scheme.

4.2.2 Data accuracy according to the probability information values link

Figure 5 shows the results of comparing the data accuracy with Shen scheme by linking the probability information values. The test results in Figure 5 show that the suggested scheme has data accuracy that is 8.3% higher than Shen scheme on average. These results can be attributed to the linkage of probability value after hierarchically constructing the data priority and attribute information by pairing them.

Figure 4. Data processing time by data probability link information.

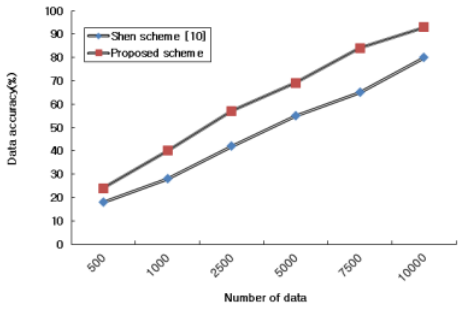

Figure 5. Data accuracy according to the probability information value link.

5. Conclusions

Search processing time and accuracy of Big data that are broadly used in diverse areas are becoming more and more important. This paper suggested a scheme that hierarchically group-manages huge volume of Big data by investigating the similarity of data using priority and attribute information. As the suggested scheme manages Big data by grouping priority information allocated to the data according to the similarity, it has faster data processing velocity than the existing scheme. The performance evaluation results showed that the suggested scheme had processing time that is 11.1% shorter and an accuracy 8.3% better than the existing scheme on average. Future study will involve the application of this study results to actual data search system and the performance evaluation of the system.

6. References

1. Hu H, Wen Y, Chua TS, Li X. Toward scalable systems for big data anaqlytics: A technology tutorial. IEEE Access. 2014 Jun; 2:652–87.
2. Russom P. Big data analytics. TDWI Research 4th Quarter; 2011.
3. Gadepally V, Kepner J. Big data dimensional analysis. Proceedings of 2014 IEEE High Performance Extreme Computing Conference (HPEC); Waltham, MA. 2014. p. 1–6.
4. Demchenko Y, Laat CD, Membrey P. Defining architecture components of the Big data Ecosystem. Proceedings of 2014 International Conference on Collaboration Technologies and Systems (CTS); Minneapolis, MN. 2014. p. 104–12.
5. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. Big data: The next frontier for innovation, competition and productivity. Mckinsey Global Institute; 2011.
6. Shen P, Zhou Y, Chen K. A probability based subnet selection method for hot event detection in sina weibo microblogging. Proceedings of 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining; Niagara Falls, ON. 2013. p. 1410–3.
7. Chen K, Zhou Y, Zha H, He J, Shen P, Yang X. Cost-effective node monitoring for online hot event detection in sina weibo. Proceedings of the 22nd International Conference on World Wide Web; NY, USA. 2013. p. 107–8.
8. Shen P, Zhou Y, Chen K. A Probability based subnet selection method for hot event detection in sina weibo microblogging. Proceedings of 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining; Niagara Falls, ON. 2013. p. 1410–3.
9. Shrivastba KMP, Rizvi MA, Singh S. Big data privacy based on differential privacy a hope for big data. Proceedings of 2014 International Conference on Computational Intelligence and Communication Networks; Bhopal. 2014. p. 776–81.
10. Shen P, Zhou Y, Chen K. A. Probability based Subnet Selection Method for Hot Event Detection in Sina Weibo Microblogging. Proceedings of 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara Falls, ON, 2013, 1410-1413.
11. Jung YC. Big data revolution and media policy issues. KISDI Premium Report; 2012.
12. Kim SH, Kim NU, Chung TM. Attribute relationship evaluation methodology for big data seucrity. Proceedings of 2013 International Conference on IT Convergence and Security (ICITCS); Macao. 2013. p. 1–4.
13. Son SY. Big data, online marketing and privacy protection. KISDI Premium Report; 2013.
14. Kim JT, Oh BJ, Park JY. Standard trends for the bigdata technologies. Electronics and Telecommunications Trends. 2013 Feb; 28(1):92–9.
15. Paryasto M, Alamsyah A, Kuspriyanto BR. Big-data security management issues. Proceedings of 2014 2nd International Conference on Information and Communication Technology (ICoICT); Bandung. 2014. p. 59–63.
16. Jeong YS, Kim YT, Park GC. Data security scheme for multiple attribute information in big data environment. Indian Journal of Science and Technology. 2015 Sep; 8(24):1–7.
17. Jeong YS. Parallel processing scheme for minimizing computational and communication cost of bioinformatics data. Indian Journal of Science and Technology. 2015 Jul; 8(15):1–8.
18. Yun SY, Min SH. A fault-tolerant bootstrap server for a system with a very large number of personal healthcare devices. Indian Journal of Science and Technology. 2015 Oct; 8(25):1–6.
19. Lee SR. Medical information security analysis for standardization strategy in Korea. Indian Journal of Science and Technology. 2015 Oct; 8(25):1–7.