Public Sentiment Big Data Query Processing and Optimization with Unified Storage of Source and Meta Data

Donglei Yan¹, Jiaxin Li², Shengnan Lei², Junri Tang², Kaiqi Kou², Zhiqiong Wang¹,*

¹College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China
²School of Computer Science and Engineering, Northeastern University, Shenyang, China

*Email: wangzq@bmie.neu.edu.cn

Abstract. Public sentiment big data has the characteristics of mass, multi-source, heterogeneity and multi-mode. At present, it mostly adopts separate storage strategies and has low storage and query efficiency. To this end, a distributed storage model is proposed in which structured and unstructured data correspond and are stored uniformly. By establishing a unified storage framework for source metadata, the original source data and characteristic metadata are stored uniformly according to different storage methods. Based on the above storage architecture, a hierarchical index structure is proposed to improve the efficiency of big data query. The unified storage model of source metadata is compared with the common storage methods, and is in the leading position in terms of Block number and query processing efficiency.

1. Introduction

With the rapid development and increasing popularity of the Internet, a wide variety of network platforms have emerged. People's lives are closely linked with the Internet. Public sentiment data in the era of big data has become a very valuable resource for public sentiment analysis business. The massive multi-source, heterogeneous and multi-modal public sentiment big data generated on different network platforms pose great challenges to data storage and query.

Effective management of public sentiment big data is an important basis to support efficient public sentiment analysis and prediction. At the same time, the query efficiency of public sentiment data also directly affects the efficiency of public sentiment analysis and prediction [1]. The traditional relational database is difficult to effectively store and query public sentiment data with the above characteristics [2]. Therefore, there is an urgent need for an effective multi-modal public sentiment data storage framework and an efficient index to manage public sentiment data.

Therefore, this paper proposes a unified storage model for multi-source heterogeneous public sentiment data, which extracts the eigenvalues of the original data and stores the source metadata in a unified way. The source data uses hierarchical key-value to organize the model files, and the metadata uses non-relational database Hbase to store. Furthermore, a hierarchical index structure combining global index and local index is adopted to realize the rapid construction of distributed index with a bottom-up strategy [3]. Based on the above ideas, this paper designs and implements a unified storage model for multi-modal source metadata, and establishes the corresponding index [4], as shown in Figure 1.
2. Storage Model and Index Design

2.1. Public Sentiment Source Data Storage and Index

2.1.1. Public Sentiment Information Model. When designing the network public sentiment data storage model, based on the characteristics of public sentiment data and HDFS [5], consider how to combine the data in an organized way to solve the problems of name node memory consumption, low query efficiency and the like when HDFS stores massive data files [6].

According to the characteristics of public sentiment information data, public sentiment information is divided into four levels of public sentiment information models. As shown in Figure 2, the first layer is the Event layer, which contains characteristic keywords that can represent the event. The second layer is the Information Source layer, which indicates the network platform where information is generated, such as microblog, Twitter, etc. The third layer is the Information layer, which is a piece of information released about an event under a certain information source and consists of one or more basic units such as text, image, audio and video. The fourth layer is the Data layer, which shows the specific content and manifestation of a certain information and is the most basic unit of the public sentiment information model. Each hierarchy has unique identifiers, namely EID, SID, IID, DID.

2.1.2. Public Sentiment Data Storage. In this paper, HDFS is used as the bottom storage system of network public sentiment big data. Based on the characteristics of public sentiment information model and HDFS's two file storage models, a two-tier storage model based on SequenceFile and MapFile is
designed. The SequenceFile organization stores all the files in the information layer, and the MapFile organization stores all the files in the information source layer in units of SequenceFile [7]. Its logical structure is shown in Figure 3.

![Diagram: Public Sentiment Data Storage Structure](image)

**Figure 3. Public Sentiment Data Storage Structure.**

In the storage model designed in this paper, the bottom SequenceFile is used as the minimum unit to store information data, and all data of various formats belonging to the same information layer are sequentially stored in the same SequenceFile file. The unique identifier DID of the data layer is taken as the Key in the SequenceFile file and the data content is taken as the Value. The < DID, data content > is serialized and stored in the Record of the SequenceFile. When the data of an event information is needed, after the SequenceFile file is queried, it can be traversed sequentially. MapFile is used as a higher-level storage unit to store public sentiment data, and all files under the information source level of the same event in the public sentiment information model are sequentially stored in the same MapFile. The unique identifier IID of the information layer is used as the Key of the record, and the sequence of SequenceFile files of the same information layer is used as the Value of MapFile records to form < IID, SequenceFile > key-value pairs and serialize and store them in MapFile. At the same time, the IID is stored in the Index of MapFile.

So far, the public sentiment information storage model designed in this paper can merge and store public sentiment information files with SequenceFile and MapFile as containers according to the public sentiment information model.

2.1.3. Multi-level Index Structure of Source Data. Based on the public sentiment data storage model, this paper designs a two-level index structure based on HBase to simulate the multi-table connection query of relational database, aiming at how to quickly and effectively query the required public sentiment information accurately in massive public sentiment data [8]. Figure 4 and Figure 5 show the first-level index structure, which consists of first-level tables and second-level tables. The first-level table stores the IID of the event, and all the information data sets of the event can be queried. According to the information release time, the information data of a certain time can be screened through the second-level table.

![Table: Level 1 Table](image)

**Figure 4. Level 1 Table.**
Figure 6 is a three-level table in the two-level index structure. After the first-level index query, the data under a certain information source can be viewed, and the data file URI of each piece of information can be viewed through the three-level table. As shown in Figure 6, the storage location of all data files for each piece of information and the Ikey of the data sequence, i.e. IID in the public sentiment information model, are stored in the column family of the HBase three-level table. Ikey is the index keyword of MapFile layer. Through the keyword, you can randomly access the files of public sentiment data to obtain the required public sentiment data, and finally realize multi-condition and efficient query of public sentiment information big data.

Figure 7. metadata storage model.
2.2.2. **Metadata Index Construction.** In the search of public sentiment information, the common search method is to find the specific content of an event through the keywords of the event. Therefore, this paper proposes to use the event keyword as RowKey to find the EID of the event. The index design model is shown in Figure 8. RowKey is the feature keyword of the event, and Text, Audio, Picture and Video column families respectively store EID of different events in each mode.

![Figure 8. metadata index structure.](image)

3. **Data Query Based on Multilevel Index**

Based on the public sentiment data storage model, the public sentiment characteristic data storage model and the corresponding index structure, this paper has built a multi-level index structure system for public sentiment big data [12].

When the query request is initiated, firstly, the corresponding event EID is found in the characteristic data index according to the keyword, then the information source is determined according to the event EID and the information release time in the first-level index of HBase, the information source is routed to the second-level index according to the SID, a certain piece of information in the information set is selected, the IKey and URI address of the information sequence are returned, and the request is sent to HDFS. HDFS locates the stored data file according to the URI address, and quickly locates the offset position of the sequence in MapFile through IKey to download and return the data.

Taking the the Belt and Road Initiative incident as an example, according to keywords such as "sharing" and "development", the event IID; is found in the metadata index. According to IID and information release time 201903110, the information source is determined to be microblog, the information Ikey and URI in the microblog are queried in the secondary index, and returned; According to the returned URI address and IKey, a request is made to HDFS, which downloads the data file using the FileSystem API and returns it to the user.

4. **Experimental Results**

4.1. **Lab Setup**

4.1.1. **Experimental Data.** The data used in the simulation experiment are multimodal data of various news collected from various information sources. At the same time, feature extraction is carried out to obtain metadata.

4.1.2. **Experiment Setup.** There are 4 hosts in Hadoop cluster under Linux environment, one of which is NameNode and the other 3 are DataNode. In the experiment, all machines belong to the same 100 MB LAN. Linux system uses open source CentOS 7.0, HDFS version 2.6. 0, and JDK version 1.8. 0 used in the system development process. The open source framework used is Dcm4che version 2.0. 29.

4.2. **Analysis of Experimental Results**

In order to fully verify the effectiveness of the public sentiment data storage model proposed in this paper to solve the problem of low efficiency in HDFS storage of massive public sentiment data files. According to the public sentiment information model, the performance test is carried out in two aspects: the number of files and the number of Blocks after merging and storing public sentiment data.

At the same time, in order to verify the effectiveness of the index structure proposed in this paper, comparative tests are carried out in two aspects: the query time when MapFile is used to store public
sentiment information, the query time directly stored in HDFS, the query time of HBase multi-level index and the query time after RowKey hashing.

![Figure 9. Comparison of the Number of Files and Blocks.](image)

As shown in Figure 9-A, after a large number of small file sets are stored according to the public sentiment information model structure, the number of files occupied in HDFS is 321, and the number of files is reduced by more than 100 times. The number of Block indexes in HDFS name nodes is significantly reduced, the memory of name nodes is saved, and the cluster work efficiency is improved.

As shown in Figure 9-B, after merging and storing small files using the public sentiment information model, the number of Blocks is reduced to 938. The decrease of the number of Blocks will reduce the number of Block indexes in NameNode, reduce the query pressure of HDFS cluster and improve the processing efficiency of HDFS cluster.

![Figure 10. Comparison of Query Time.](image)

Figure 10-A shows the comparison between the query time using HBase index after storing public sentiment data according to the storage and index designed in this paper and the query time directly storing public sentiment data in HDFS. HBase index significantly reduces the query time of information files. Therefore, the use of HBase index can improve the efficiency of public sentiment information big data query.

Figure 10-B compares the HBase multilevel index with the query time after RowKey hashes with MD5. After RowKey is hashed by MD5 algorithm, the problem of "data tilt" of HBase is well solved, and the efficiency is obviously improved when querying. At the same time, the effectiveness of hashing RowKey of HBase based on MD5 algorithm proposed in this paper is proved.

5. Conclusion
In this paper, a large number of small file sets are stored according to the public sentiment information model structure, and the number of files is greatly reduced. After merging and storing small files, the number of Blocks is significantly reduced, which improves the work efficiency of HDFS cluster and can quickly index relevant information. At the same time, it can be seen from experiments that the query efficiency of HBase multi-level index hashed by MD5 is significantly better than other index structures. The validity of the index structure proposed in this paper is fully verified. However, this method is only suitable for data with layered characteristics, and due to layered merging of files, the reading and writing efficiency is slightly reduced, which needs further improvement and is applied to mature database systems.

Acknowledgments
This research was partially supported by the CETC Joint Fund, the National Natural Science Foundation of China (No. 62072089), the China Postdoctoral Science Foundation (Nos. 2019T120216 and 2018M641705), the Fundamental Research Funds for the Central Universities (Nos. N2019007, N180101028, N180408019 and N2024005-2) and the fund of Acoustics Science and Technology Laboratory.

References
[1] Wang J, Wu S, Gao H, Li J and Ooi B C 2010 Management of Data
[2] Fan W and Huai J-P 2014 Journal of Computer Science and Technology 29 21
[3] Doulkeridis C and Nørvåg K 2014 The VLDB Journal 23 26
[4] Ahad M A and Biswas R 2018 Procedia Computer Science 132 10
[5] Cai B and Li W 2019 Proceedings of International Conference on AI and Big Data Application P 5
[6] Wang X, Zhao D, Yang M, Duan L, Xiang M M and Guo Q 2017 Information Discovery and Delivery 45 14
[7] Chen G, Nie P and Jing W 2018 International Journal of Performability Engineering 14 2545
[8] Tao Y, Lin W and Xiao X 2013 Management of Data
[9] Maliberan E V, Sison A M and Medina R P 2018 International Journal of Communication Networks and Information Security 10 374
[10] Zhang F, Song Y, Cai W, Liu S, Liu S, Pujol S, Kikinis R, Xia Y, Fulham M J, Feng D D Alzheimers Disease Neuroimaging Initiative J 2016 IEEE Transactions on Bio-Medical Engineering 63 12
[11] IDERA J 2019 Information Technology Newsweekly
[12] Wang J and Duan S 2019 Proceedings of 2019 3rd International Conference on Informatization in Education, Management and Business P 5