ATLANTIS PRESS

Advances in Engineering Research (AER), volume 143

6th International Conference on Energy and Environmental Protection (ICEEP 2017)

Analysis and Establishment of Big data Model

Yachuan Yao[1, a], Qiang Han[2,b], Yongchun Liu[3,c]

[1, 2, 3] School of Physics & Electronic Engineering, Sichuan University of Science & Engineering, Zigong, SiChuan, 643000, China

[a]610851229@qq.com, [b]hanqiang770@126.com, [c]53571176@qq.com

Keywords: Big data, Unstructured data, Data model, Retrieval analysis, visualization

Abstract. Big data service is a new data resource model and a new service economy model, which encapsulate all kinds of big data operations, to provide consumers with services everywhere, standardization, on-demand retrieval, analysis and visualization service. An unstructured data model based on agent behavior is designed in this paper。It describes the process of retrieval, analysis and visualization services, and improve the accuracy and efficiency of the service retrieval service two measures to implement data service capacity optimization in the big data service application.

Introduction

The existing study of data services can not be applied to Big data services. Big data services are faced with the following challenges[1]:

(1) Data sources in Big data environments are not only structured data, but more unstructured data. In order to enable Big data services to support unstructured data, it is necessary to design a common data model that can express unstructured basic information and background information, thus laying the foundation for the establishment of Big data services for unstructured data.

(2) The major requirements for Big data are search, analysis and visualization. The problem that Big data services need to solve is how to provided search, analysis and visualization to consumers in the form of service. In addition, in order to adapt to the changing service scenarios and personalized user needs, it is also a facing challenge that how to improve the services ability of Big data.

This paper mainly studies in the following aspects.

(1) Aiming at the problem of lacking unstructured data model which is suitable for Big data service, we design a general data model covering multiple attributes such as unstructured data basic information, domain information and task information. At last, examples are verified by lightweight description language and operating language.

(2) Research on the processing flow of keyword search requests and semantic retrieval request, and the processing of strategy and implementation for analysis and visualization of the request for Big data services is made through in-depth analysis of Big data retrieval, analysis and visualization requirements. Practical projects are made to verify the case[2]. At the same time, in order to optimize the ability of Big data services, study of the search results ranking optimization algorithm and persistent cache mechanism based on mixed prefetch are carried out respectively from the two aspects that improve the accuracy of search results and service efficiency.

An unstructured data model based on subject behavior

Description of requirement. In this paper, galaxy data model (GDM) based on the behavior of the main body of the unstructured data is proposed by analyzing the behavior characteristics of the data manipulation subject and considering the external factors such as the background of the data generation and the domain.

In addition to the various types of data, the unstructured data has obvious user characteristics, many kinds of storage media, diverse applications and other characteristics. Among them, the user characteristics are the core features of unstructured data, so we are user-oriented when design unstructured data model. The unstructured data model is an abstract concept that primarily addresses how to express unstructured data.

GDM defines the concepts of the subject, the data object, the attribute, the attribute class, the attribute type and the galaxy model by introducing the concept of the BianTi, and describes the information related to the unstructured data in the multidimensional data space.

Furthermore, for Galaxy models have richer data attributes, the unstructured data management system based on the GDM model is able to meet the complex retrieval requirements, such as supporting the retrieval of "all the information on the project of Big data management key technology research", in addition to supporting the keyword search.

Model implementation and examples. In Big data environment, unstructured data has real-time and massive features. On the one hand, unstructured data is generated in real time, and its data properties are characterized by evolution with various factors such as time, event, subject and so on. Therefore, scalability is very important when implementing galaxy model[3]. On the other hand, unstructured data is massive, attribute information as a subsidiary of massive unstructured data is required to occupy as little space as possible. Therefore, Javascript Object Notation (JSON) is used in this paper to describe the galaxy model. It is a lightweight data exchange format, which has the advantages of lightweight, scalable, and fast indexing, and is suitable for describing unstructured data.

Example of the data model construction. This section constructs a data model for a document -type data "water analysis.docx" and provides an example of the implementation of two complex data searches.

Each attribute class of the document type data is described, extracted and set the attribute value. The attribute class of the data is described in terms of basic attributes, content attributes, feature attributes, task attributes, and environment attributes,such as:

(1)basic attribute

“BasicAttr”:{
“FileAttr”:{“Size”:”528KB”,”CreatTime”:”2017-3-1 09:15”,”ModifyTime”:”2017-3-6 10:15”
“FolderPath”:”hdfs://waterpollution/”,”FileType”:”Microsoft Office Word”, ”Name”:”Water Pollution Monitoring”},
“SourceAttr”:{“Author”:”Han Mei”,”Program”:”Microsoft Office Word Document”,”CreateContentTime”:” 2017-3-1 09:15”,”LastSave Time”:” 2017-3-6 10:15”},
“AuthorityAttr”:{“AuthorityType”:”FullControl”,”Steward”:[“Song nana”,”Jing”], ”Company”:”BUPT”}}

(2)content attribute

“ContentAttr”:{
“DescriptionAttr”:{“Title”:” Water Pollution Monitoring”,
“Topic”:”key technology of data management”:”Language”:[“English”,”Chinese”]},
“SemanticAttr”:{“Tag”:[“water pollution”,”monitoring”,”Fuxi River”],”Field”:” Environmental Science”,”URI”:file://D:\\Project\\Water}}

Data retrieval example. JAQL can operate directly on data stored in HDFS. To achieve parallelism, JAQL can also rewrite high-level queries as low-level queries consisting of MapReduce jobs at the appropriate time. The engine's internal conversion of queries and tasks can significantly reduce application development time associated with analyzing large amounts of data in Hadoop. Therefore, JSON is used to represent and store unstructured data model in this paper. It uses the JAQL to descript the query operation, so as to eventually achieve the massive unstructured data retrieval in Big data.

Taking the Big data research in the "Networked Data Resource Management" project as an example, Undata.json is a json file that describes the galaxy data model. This paper presents an example of unstructured data retrieval based on GDM modeling through the following two search scenarios.

Search statements:

$undata=read(hdfs(“UnData.json”))//
->filter $.BehaviourAttr.TaskAttr.TaskName LIKE”%big data management %”
Or $.BehaviourAttr.TaskAttr.TaskName LIKE”%big data management %”

->group by type=$.BasicAttr.fileAttr.filetype into {type,name:$.BaiceAttr.Name,path:
$.BasicAttr.FileAttr.FolderPath};
Search results:
[{
“type”:”Microsoft Office Word document”,
“name”:”Minutes of Big data management seinar”,
“path”:”F:\\Project\\unstructured data”},
{“type”:”AVI Video”,
“name”:”Video of Big data management seminar”,
“path”:”F:\\Project\\Unstructured data”
}]

Research on Application of Big data Services

An optimization algorithm for unstructured data retrieval ranking and HPPC based on hybrid prefetch and persistent cache mechanism are proposed to improve the accuracy and service efficiency of retrieval results, so as to optimize the data service capability[4]

Existing problems. In the aspect of data retrieval, there are some problems in the existing retrieval of data services, such as lack of support for keyword retrieval, low accuracy of unstructured data retrieval results, and no support for unstructured data retrieval. In the aspect of data analysis, the main service method is that the data owners provide data, analysis technology providers use their own data analyst, analysis products and computing resources to help users complete data analysis, which brings a very high cost to the user and also great limits to the development of Big data analysis. In terms of visualization, the existing visualization is achieved in a manner of professional systems or programs developed by professionals. That means there is a high threshold like data analysis, which largely restricts the insight into data laws through visualization.

Data retrieval services. The current research on the ranking of unstructured data retrieval mainly focuses on the data itself, and does not take into account the impact of the relationship between data and user behavior on search rankings. At the same time, existing research involves less data importance. Some documents refer to the different importance of files, but did not give a specific algorithm.

Therefore, the Big data service architecture needs to support two types of search modes.

Semantic Search: A professional or an application retrieves a search for a search service.

Keyword Retrieval: The user submits a search request in the form of a search keyword or an "attribute value" to start search,it is shown in table 1.

Table 1 Keyword Retrieval example

Retrieval Type	Meaning
‘K1’=’V1’;S=’WaterPollution’	In data services " WaterPollution ",retrieve all data whose K1 attribute value isV1
‘K2’=’V2’ or ‘K3’=’V3’;T=R	In all structured data source services, the K2 property is retrieved with V2, or the data of the K3 attribute is V3
‘K1’=’V1’ and ‘K4’=’V4’;T=U;	In all unstructured data source services, retrieve the data with the K1 attribute value of V1 and the K4 attribute value of V4
Standard Specification for river water quality analysis	Retrieves all the data that contains the keyword

Data analysis services.

The processing flow of analysis request for large data services is as follows[5]:

Step1 The user generates a data analysis request including the target data source, the analysis rule set and the analysis result.

Step2 The service matching component locates the required data analysis service from the data service registry according to the data source entered in the analysis request.

Step3 The service composition component generates the data service composition and the result assembly rule according to the user request.

Step4 Associating with the request analysis component, the analysis request is decomposed into a series of S-oriented sub-analysis request.

Step5 The rule decomposition component decomposes the analysis rule set into a plurality of disjoint subsets corresponding to each sub-analysis request.

Step6 Sub-analysis request and the rule subset are dispatched to S respectively.

Step7 S performs the analysis task in parallel, and obtains the temporary analysis result R respectively.

Step8 Assemble the R as required according to the rule.

Step9 Output R.

Data visualization services.

The processing flow of visualization request for large data services is as follows:

Step1 Accept the user's retrieval request Q_S or analysis request Q_A, and the user-supplied visualize graphics G (optional) and the output script type T.

Step2 Execute Q_S and Q_A according to the large data retrieval service and analysis service execution process and obtain the temporary result set R_S^* and.

Step3 The data characteristics of R_S^*or R_A^* are analyzed to confirm whether the data feature matches the graph G, and if not, we need to select the appropriate visual graph G^*.

Step4 The service matching component selects a matching visualized data service S_1 in the Data Services registry.

Step5 Enter R_S^* or R_A^* into S_1 to perform the visualization task and output the visual script based on type T.

Experiment and result analysis .

The basic idea of this experiment is to get the original search results which is the experimental data also, and then use the algorithm to rank the results[6].

Through the 14-day monitoring of the large data service retrieval operation of the experimenter, all operations, search terms, search results and sorting of unstructured data were recorded in the report. And after the monitoring, the experimenter described the purpose of each search, identify the relevance of the search results, and improve the search results.

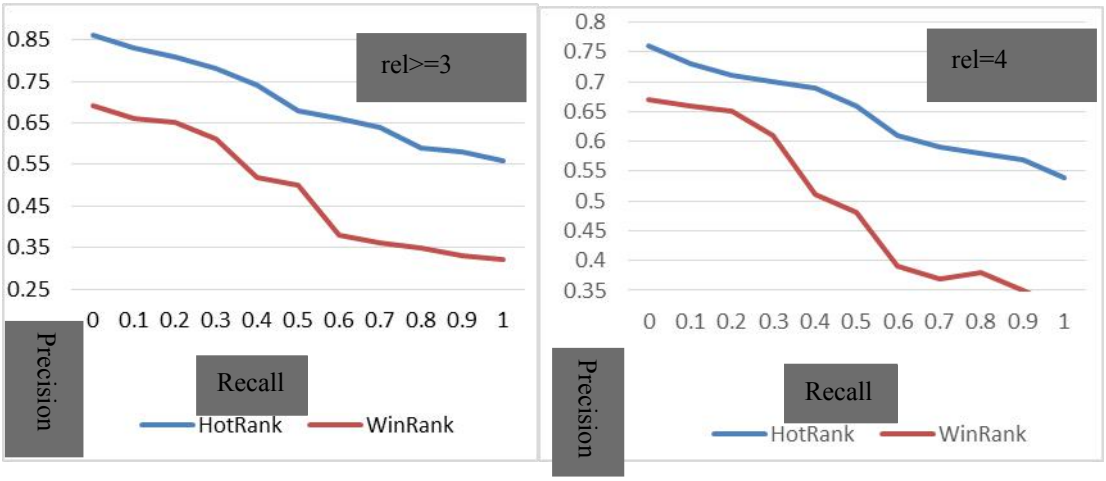

Fig.1 11point P-R curve

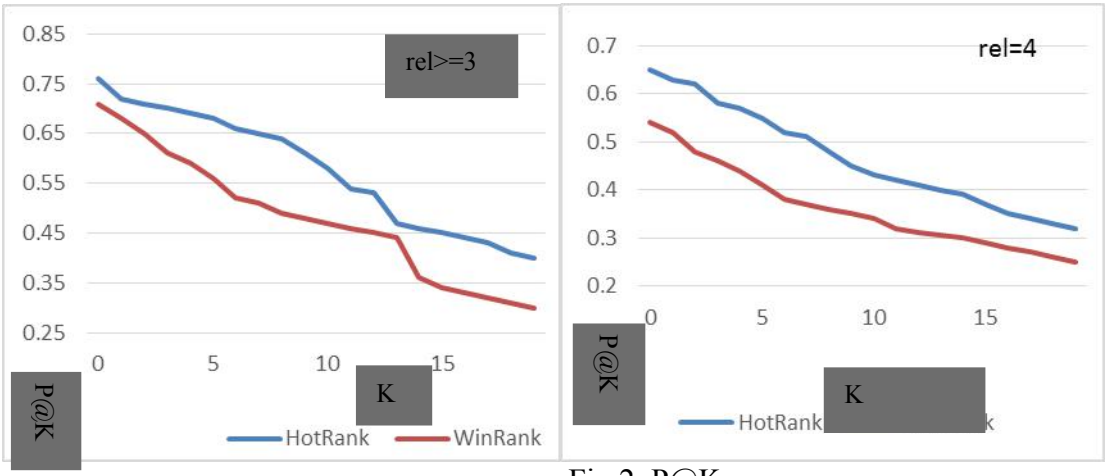

Fig.2 P@K

Service response optimization based on mixed prefetch and persistent cache .

Requirement description

It meet the relational database storage and unstructured data storage of the dual requirements, and have storage management capabilities for massive super image, log data and other unstructured data.It has a faster response speed for complex queries and high concurrent connection requests.It has high scalability.

Architecture design. A hybrid prefetch based Persistent Caching (HPPC) is designed in this paper, shown in figure 3.

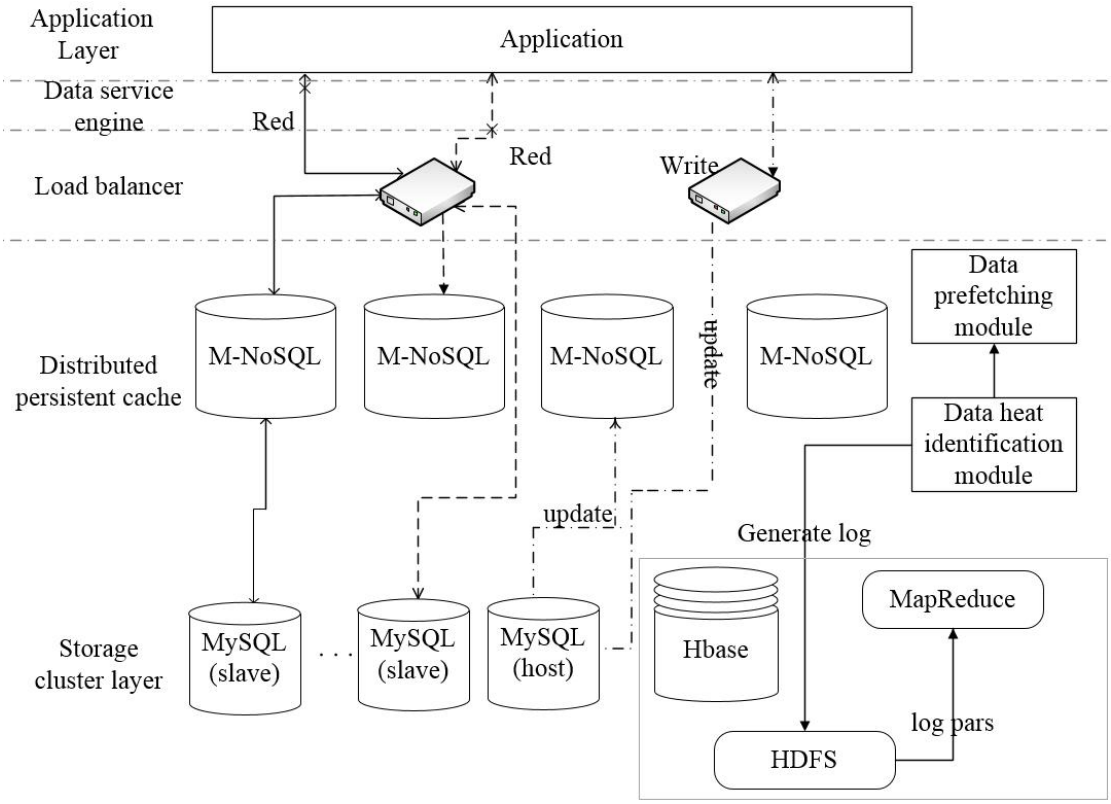

Fig.3 Hybrid Prefetch based Persistent Caching，HPPC

Hybrid Prefetch Algorithm and persistent cache. In cache layer design, we can meet the following two features: cache is not limited by memory and is not easily lost; it has the ability to dynamically prefetch hot spot data.

A hybrid prefetch algorithm based on data heat recognition has been presented in this paper, as shown below[7].

```
Required: SystemLog SL
Begin
   for all sl ∈SL do
    Score←Calculate(sl)
   End for
   for all sl ∈ Score do
    If sc not NULL then
     Rule=Transform(sc)
     If rule blongs to static rule then
      staticRuleSet.append(rule)
     else dynamicruleSet.append(rule)
     End if
    End if
   End for
   get CandidateData according RuleSet
Return CandidateData
```

Based on the above algorithms, the actual flow of data prefetching is shown in Figure 4:

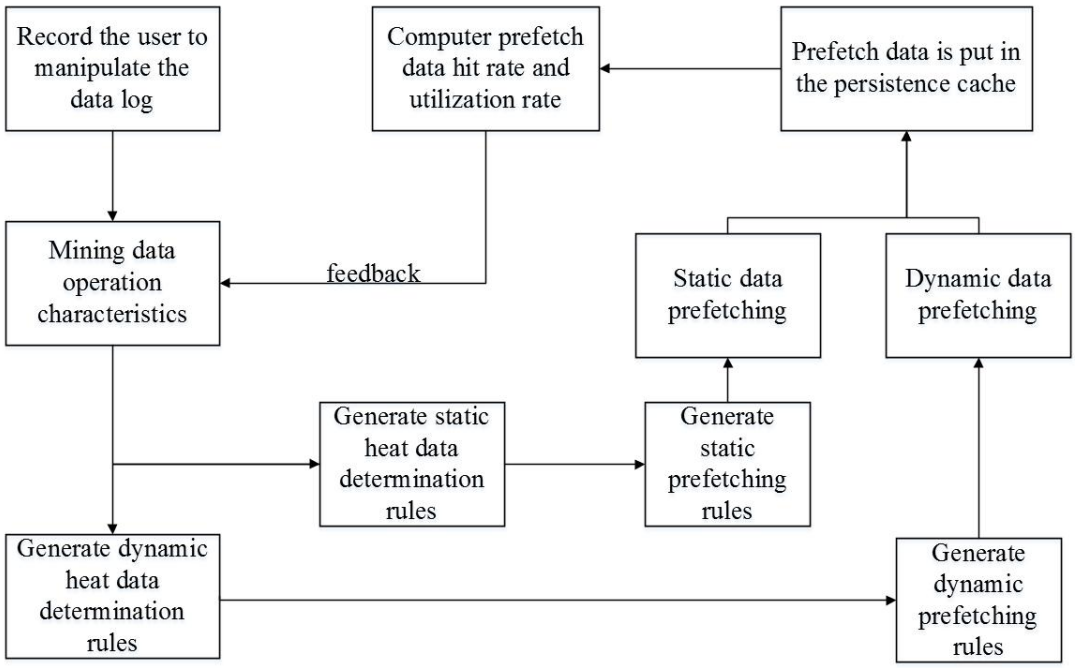

Fig.4 Hybrid Prefetching Process Based on Data Thermal Recognition

Experiment and result analysis. The hybrid prefetching algorithm based on data heat identification was applied to the project of river water quality monitoring, and the operation behavior and prefetch execution were recorded in 1 month period. To evaluate the performance of the mixing prefetching algorithm by selecting 7 operating statements which were more frequently used. Hit rate and Prefetch accuracy used evaluation indicators were selected from the evaluation criteria of the prefetching algorithm performance. The experimental data are shown in Table 2.

Table 2 Experimental results analysis of HPA algorithm

Operation	R	P	P+	HR(%)	Precision(%)
1#	3744	4215	2593	69	62
2#	3588	4849	2193	61	45
3#	3267	4682	1405	43	30
4#	2091	2363	1281	61	54
5#	1421	1811	812	57	45
6#	1254	1567	426	34	27
7#	755	931	352	47	38
Overall evaluation	12160	20418	9062	53	45

The experimental results show that the average hit rate of HPA prefetching is 53%, and the average precision is 45%, which shows that the algorithm has good ability of user operation, data prediction and optimization.

Conclusions

The paper proposed a Galaxy Data Model of unstructured data based on agent behavior, proposed A heat sensitive unstructured data retrieval ranking optimization algorithm(HotRank), proposed a Hybrid Prefetch Algorithm based on data heat recognition. The experiment shows that the expected effect has been achieved.

Acknowledgements

This work was financially supported by Enterprise information and Internet of things measurement and control technology of Key Laboratory of Sichuan province open project (2014WYJ05), Zigong science and Technology Bureau project(2014DZ10),project of Liquor making biological technology and application of key laboratory of Sichuan province(NJ2014-14), Major training project of Sichuan University of Science and Engineering(2014PY15)

References

[1] Doan A,Naughton J F,Baid A,etal. The case for a structured approach to managing usstructured data[C]. In:Proceedings of the Fourth Biennial Conference on Innovation Data System Research.Asilomar,2009

[2] Li W,Lang B.A tetrahedral data model for unstructured data management[J].Science China Information Sciences,2010,53(8): p. 1497-1510.

[3] Crockford D. Introducing json [J].Available:json.org,2011.

[4] 1010data[EB/OL] http://www.1010data.com.

[5] Redis[EB/OL] http://redis.io

[6] Baowen Xu,Weifeng Zhang. Application of data mining technology in WEB prefetch [J].Chinese Journal of Computer ,2001,24(4): p. 430-436

[7] Jin Han,Meila Song.Big data 2[J].ZTE TECHNOLOGY JOURNAL.2013,4